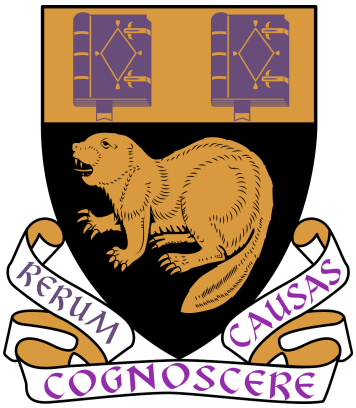
Eigenvalue-Regularized Covariance Matrix Estimators for High-Dimensional Data



Huang Feng

Department of Statistics

London School of Economics and Political Science

A thesis submitted for the degree of

Doctor of Philosophy

November 2018

致我最爱的家人

To my loving family.

Declaration

I hereby declare that the thesis submitted for the PhD degree of the London School of Economics and Political Science is my own work except where specific reference is made to the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it). This thesis has not previously been submitted in whole or in part for consideration for any other degree.

I confirm that Chapter 3, 4 and 5 are jointly co-authored with Dr. Clifford Lam.

Huang Feng
November 2018

Acknowledgements

First of all, I would like to thank my supervisor, Dr. Clifford Lam, for his immense support, inspiring guidance and invaluable encouragement throughout my PhD study. I am extremely lucky to have a supervisor who cares so much about my work. I could not ask for more from such a nice supervisor. I would also like to show my deep gratitude to my second supervisor, Professor Piotr Fryzlewicz, for his suggestion and recommendation for my pursue of PhD study.

I would like to express my great appreciation to all the staff and colleagues in the Department of Statistics at the LSE, for making the last four years a great experience. Special thanks to all my lovely officemates Andy Ho, Haziq Jamil, Tayfun Terzi, Yajing Zhu, Ragvir Sabharwal, Yan Qu, Xiaolin Zhu, Alice Pignatelli di Cerchiara and Filippo Pellegrino for helping me with my endless academic and non-academic questions and giving me great support for surviving the PhD. I wish to acknowledge the help provided by the students of my supervisor Charlie Hu and Cheng Qian for the efficient discussions about the research. I am particularly grateful for the invitation from Dr. Wicher Bergsma, Dr. Matteo Barigozzi and Dr. Yining Chen to be their teaching assistant, and same thanks to my teaching partner José Manuel Pedraza Ramirez for making wonderful teaching experience as a team. Assistance provided by Ian Marshall and Penny Montague was greatly appreciated. Besides, my research would not have been possible without the sponsorship from LSE.

Finally, I wish to thank my dearest parents and grandparents for their boundless love and consistent support, especially my grandfather for his open mind and great encouragement, and also my father for being my first teacher to initiate me into mathematics. My special thanks are extended to Xue Lu, Hao Liu, Longjie Jia, and Luting Li, who are more than friends like family to me in London, for always being there for me no matter what. Further thanks to all my loving friends and relatives here in UK, back in China, and elsewhere in the world to whom I am indebted for their constant encouragement.

Abstract

Covariance regularization is important when the dimension p of a covariance matrix is close to or even larger than the sample size n. This thesis concerns estimating large covariance matrix in both low and high frequency setting.

First, we introduce an integration covariance matrix estimator which is a linear combination of a rotation-equivariant and a regularized covariance matrix estimator that assumed a specific structure for true covariance $\mathbf{\Sigma}_0$, under the practical scenario where one is not 100% certain of which regularization method to use. We estimate the weights in the linear combination and show that they asymptotically go to the true underlying weights. To generalize, we can put two regularized estimators into the linear combination, each assumes a specific structure for $\mathbf{\Sigma}_0$. Our estimated weights can then be shown to go to the true weights too, and if one regularized estimator is converging to $\mathbf{\Sigma}_0$ in the spectral norm, the corresponding weight then tends to 1 and others tend to 0 asymptotically. We demonstrate the performance of our estimator when compared to other state-of-the-art estimators through extensive simulation studies and a real data analysis.

Next, in high-frequency setting with non-synchronous trading and contamination of microstructure noise, we propose a Nonparametrically Eigenvalue-Regularized Integrated coVariance matrix Estimator (NERIVE) which does not assume specific structures for the underlying integrated covariance matrix. We show that NERIVE is positive definite in probability, with extreme eigenvalues shrunk nonlinearly under the high dimensional framework $p/n \to c > 0$. We also prove that in portfolio allocation, the minimum variance optimal weight vector constructed using NERIVE has maximum exposure and actual risk upper bounds of order $p^{-1/2}$. The practical performance of NERIVE is illustrated by comparing to the usual two-scale realized covariance matrix as well as some other nonparametric alternatives using different simulation settings and a real data set.

Last, another nonlinear shrinkage estimator of large integrated covariance matrix in high-frequency setting is explored, which shrinks the extreme eigenvalues of a realized covariance matrix back to an acceptable level, and enjoys a certain asymptotic efficiency when the number of assets is of the same order as the number of data points. Novel maximum exposure and actual risk bounds are derived when our estimator is used in constructing the minimum-variance portfolio. In simulations and a real data analysis, our estimator performs favourably in comparison with other methods.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Estimation of a covariance matrix or its inverse, called the precision matrix, is an important and sometimes inevitable task in data analysis. Examples range from risk estimation and portfolio allocation in finance to classification or multiple hypotheses testing in general statistical analysis. Although it is easier than ever nowadays to obtain relatively large data set for study, such richness of data also means that more often than not the data we obtain are high-dimensional in nature, in the sense that the number of variables under study is comparable to or even larger than the sample size. This creates problems for traditional estimator such as the sample covariance matrix. Well documented in Bai and Yin (1993) and subsequent random matrix theory researches (see Bai and Silverstein (2010), for example), a particularly serious problem is that the eigenvalues of the sample covariance matrix are more extreme than their population counterpart. Moreover, when the dimension p is larger than the sample size n, the sample covariance matrix is not invertible, where regularization is needed.

This thesis contains three parts where covariance matrix estimation under different data frequency settings is applied: weighted average of a rotation-equivariant and a regularized covariance matrix estimator under low-frequency setting, a nonparametrically eigenvalue-regularized integrated covariance matrix estimator and a nonlinear shrinkage estimator of large integrated covariance matrix under high-frequency setting. The thesis is organised as follows.

In Chapter 2, state-of-the-art regularization methods developed for covariance matrix estimation are reviewed. One main branch of the estimations assumes a special structure of the population covariance matrix $\mathbf{\Sigma}_0$ or the corresponding precision matrix $\mathbf{\Sigma}_0^{-1}$. The commonly exploit structure in applications include the sparseness of $\mathbf{\Sigma}_0$

(Bickel and Levina, 2008a; Cai and Liu, 2011; Rothman et al., 2009), bandable type structure (Bickel and Levina, 2008b; Cai and Yuan, 2012; Cai et al., 2010), and a factor structure (Fan et al., 2008, 2011, 2013). Another branch concerns with regularizing, or "shrinking", the eigenvalues of the sample covariance matrix under the high-dimensional setting $p/n \to c > 0$. Research include the linear shrinkage (Ledoit and Wolf, 2004) and nonlinear shrinkage (Abadir et al., 2014; Huang and Fryzlewicz, 2015; Lam, 2016; Ledoit and Wolf, 2012, 2013).

While both branch of researches provide good estimators under different scenarios, there are no methods that can "take advantage" of what each branch of estimators can offer. In Chapter 3, we introduce an integration covariance matrix estimator that is a linear combination of a rotation-equivariant and a regularized covariance matrix estimator that assumed a specific structure for $\mathbf{\Sigma}_0$, under the practical scenario where one is not 100% certain of which regularization method to use. By minimizing the Frobenius loss, we derive explicit formulae for the weights which can be estimated consistently, or even almost surely, through a data splitting scheme which is similar to the one in Lam (2016). To generalize, we can put two regularized estimators into the linear combination, each assumes a specific structure for $\mathbf{\Sigma}_0$. Our estimated weights can then be shown to go to the true weights, and if one regularized estimator is converging to $\mathbf{\Sigma}_0$ in the spectral norm, the corresponding weight then tends to 1 and others tend to 0 asymptotically. Extensive simulations also reveal that our estimator can indeed gather the advantages from a regularized estimator and perform well, even when the regularized estimator itself does not. We also show that our estimator is asymptotically efficient when compared to an ideal estimator constructed with the knowledge of $\mathbf{\Sigma}_0$.

In Chapter 4, we propose a nonparametrically eigenvalue-regularized integrated covariance matrix estimator (NERIVE) which does not assume a specific structure for the underlying integrated covariance matrix in high-frequency data analysis. Under such setting, the extreme eigenvalues of a realized covariance matrix are biased when its dimension p is large relative to the sample size n. Together with non-synchronous trading and contamination of microstructure noise, the associated challenges have to be overcome at the same time. By a data splitting method, we show that the resulting integrated covariance matrix estimator is consistent with a certain positive definite matrix with regularized eigenvalues at a rate of $n^{-1/6}$ under the setting $p/n \to c > 0$.We also prove that in portfolio allocation, the minimum variance optimal weight vector constructed using NERIVE has maximum exposure and actual risk upper bounds of order $p^{-1/2}$. Incidentally, the same maximum exposure bound is also satisfied by

the theoretical minimum variance portfolio weights. All these results hold true also under a jump-diffusion model for the log-price processes with jumps removed using the wavelet method proposed in Fan and Wang (2007). They are further extended to accommodate the existence of pervasive factors such as a market factor under the setting $p/n \to c > 0$. The practical performance of NERIVE is tested by a traditional portfolio allocation problem, where our estimator is compared to other state-of-the-art estimators including the usual two-scale realized covariance matrix.

Finally in Chapter 5, we modify the data splitting method similar to NERIVE to achieve nonlinear shrinkage of eigenvalues in a covariance matrix. It produces a positive definite estimator of the integrated covariance matrix asymptotically almost surely, and involves only eigen-decompositions of matrices of size $p \times p$, which are not computationally expensive when p is of the order of hundreds, the typical order in portfolio allocation. We also present the maximum exposure and actual risk bounds for minimum variance portfolio construction using our estimator. The maximum exposure bound is of particular importance, as it is shared by the theoretical minimum-variance portfolio which assumes that the integrated covariance matrix is known. The practical performance is compared to other popular estimators.

Chapter 2

Review on Covariance Matrix Estimation

In this chapter, we provide a review on the estimation of covariance matrix or its inverse (known as precision matrix) related to this thesis, as well as the most popular state-of-the-art covariance estimators in applications for both high-dimensional and high-frequency data setting.

The most commonly used estimator is the sample covariance. Let $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, where $\mathbf{y}_t = (y_{1t}, y_{2t}, \ldots, y_{pt})^{\mathrm{T}}$'s are independent $p \times 1$ vectors for $t = 1, \ldots, n$, being the data observed at time t, for example, the stock market daily returns. Let $\mathbf{\Sigma}_0 = (\sigma_{ij})_{p\times p}$ be the population covariance matrix. The $p \times p$ sample covariance matrix is defined as

$$\widehat{\mathbf{\Sigma}}_{\text{sam}} = \frac{1}{n-1}\sum_{t=1}^{n}(\mathbf{y}_t - \bar{\mathbf{y}})(\mathbf{y}_t - \bar{\mathbf{y}})^{\mathrm{T}},$$

where $\bar{\mathbf{y}} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_i$ is the sample mean. $\widehat{\mathbf{\Sigma}}_{\text{sam}}$ is an unbiased estimator and easy to calculate. We can also use the maximum likelihood estimation (MLE) of the covariance matrix

$$\widehat{\mathbf{\Sigma}}_{\text{MLE}} = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{y}_t - \bar{\mathbf{y}})(\mathbf{y}_t - \bar{\mathbf{y}})^{\mathrm{T}}.$$

for the Gaussian distributed data. The use of the coefficient $1/n$ instead of $1/(n-1)$ makes $\widehat{\mathbf{\Sigma}}_{\text{MLE}}$ a biased estimator. The ratio of $1/(n-1)$ to $1/n$ tends to 1 when n is sufficiently large, however, making this MLE covariance approximately equal to the sample covariance $\widehat{\mathbf{\Sigma}}_{\text{sam}}$.

We enjoy the simplicity of the calculation of $\widehat{\mathbf{\Sigma}}_{\text{sam}}$ and $\widehat{\mathbf{\Sigma}}_{\text{MLE}}$, however, they are well known to have poor performance when the dimension p is larger than or comparable to the sample size n. In this high-dimensional setting, sample covariance matrix is no longer a consistent estimator as the eigenvalues of $\widehat{\mathbf{\Sigma}}_{\text{sam}}$ are not converging to the true ones based on the random matrix theory (Chen et al., 2013; Johnstone, 2001; Marčenko and Pastur, 1967). Also, its inverse, the sample precision matrix $\widehat{\mathbf{\Sigma}}_{\text{sam}}^{-1}$, is not defined due to the singular sample covariance matrix.

As a result, regularization is needed in high-dimensional covariance matrix estimation. There are two main branches for estimating a large covariance matrix nowadays: one assumes a specific structure of the population covariance matrix, sparsity as an example; the other branch focuses on regularizing, or we say 'shrinking', the eigenvalues of the sample covariance matrix. Here, we start from the sparsity setting of the large covariance estimation.

2.1 Estimation of Covariance Matrix through Thresholding

2.1.1 Simple Thresholding

To deal with the ill-conditioned estimation problem of sample covariance matrix in high-dimensional setting, thresholding is one of the simplest method to apply under the assumption that the population covariance matrix is sparse or approximately sparse, i.e. most of the non-diagonal elements in the matrix are zero or nearly but not exactly zero (Bickel and Levina, 2008a). The uniformity class of the approximately sparse covariance matrices is defined as

$$\mathcal{U}(q, c_0(p), M) = \{\mathbf{\Sigma} : \sigma_{ii} \leq M, \max_i \sum_{j=1}^{p} |\sigma_{ij}|^q \leq c_0(p)\}, \tag{2.1}$$

for $0 \leq q < 1$. If $q = 0$, the matrix is truly sparse; otherwise $0 < q < 1$, the matrix is approximately sparse. According to Bickel and Levina (2008a), if some elements of the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\text{sam}} = (\widehat{\sigma}_{ij})_{p\times p}$ have small values, it can be thresholded to a

new estimator $\widehat{\mathbf{\Sigma}}_{\text{thre}} = (\widetilde{\sigma}_{ij})_{p\times p}$ defined as

$$\widetilde{\sigma}_{ij} = \begin{cases} \widehat{\sigma}_{ij} & \text{for } i = j, \\ \widehat{\sigma}_{ij}\mathbb{1}_{\{|\widehat{\sigma}_{ij}|>\omega\}} & \text{for } i \neq j, \end{cases} \tag{2.2}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function and ω is the threshold parameter. This is called the hard thresholding. They proved the consistency of this estimator if the variables $\mathbf{y}_i$'s are Gaussian or sub-Gaussian, uniformly on $\mathcal{U}(q, c_0(p), M)$, $\log p/n = o(1)$, and $\omega = C(\log p/n)^{1/2}$ for sufficiently large C,

$$\|\widehat{\mathbf{\Sigma}}_{\text{thre}} - \mathbf{\Sigma}_0\| = O_p(c_0(p)\omega^{1-q}), \qquad \|\widehat{\mathbf{\Sigma}}^{-1}_{\text{thre}} - \mathbf{\Sigma}_0^{-1}\| = O_p(c_0(p)\omega^{1-q}),$$

where $\|\cdot\|$ is the operator norm, also known as spectral norm or l_2 matrix norm, as $\|\mathbf{A}\| = \lambda^{1/2}_{\max}(\mathbf{A}^{\mathrm{T}}\mathbf{A})$ for any matrix $\mathbf{A}$. The choice of threshold ω can be estimated by cross-validation method.

Thresholding decreases the estimation errors for small valued elements, since the errors are not accumulated with estimation. On the other hand, choosing elements that should be thresholded is always easier than estimating these small values. Although positive definiteness of $\widehat{\mathbf{\Sigma}}_{\text{thre}}$ cannot be gaurenteed, the probability of it being positive definite is approaching to 1 as long as $\log p/n \to 0$. Despite on the simplicity and computational advantage, simple thresholding neglects the difference of covariance scales, i.e. the observed data series are usually on different scales. To solve the scale inconsistency problem, Cai and Liu (2011) proposed the entry-dependent adaptive thresholding, taking varying scales into account.

2.1.2 Adaptive Thresholding

One of the most natural and simplest solutions to deal with the variability of variances is to threshold on the sample correlation matrix instead of the sample covariance matrix. Denote the population correlation matrix $\mathbf{R}_0 = (r_{ij})_{p\times p}$ and sample correlation matrix $\widehat{\mathbf{R}}_{\text{sam}} = (\widehat{r}_{ij})_{p\times p}$, where $\widehat{r}_{ij} = \widehat{\sigma}_{ij}/\sqrt{\widehat{\sigma}_{ii}\widehat{\sigma}_{jj}}$. The thresholded correlation matrix estimator $\widehat{\mathbf{R}}_{\text{thre}} = (\widetilde{r}_{ij})_{p\times p}$ is defined by

$$\widetilde{r}_{ij} = \begin{cases} \widehat{r}_{ij} & \text{for } i = j, \\ \widehat{r}_{ij}\mathbb{1}_{\{|\widehat{r}_{ij}|>\omega\}} & \text{for } i \neq j, \end{cases}$$

and the corresponding covariance matrix is $\widehat{\mathbf{\Sigma}}_{\mathrm{R}} = \mathbf{D}^{1/2}\widehat{\mathbf{R}}_{\mathrm{thre}}\mathbf{D}^{1/2}$, where $\mathbf{D}$ is the diagnal matrix of the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\mathrm{sam}}$. It is easily to see that $\omega = C\sqrt{\log p/n}$ is a good threshold choice for any constant $C > 0$, and if C is sufficiently large, it can be shown that minimax rate of convergence is $c_0(p)\omega^{1-q}$, same as the simple thresholding above. This method well solves the variability of variance, nevertheless, the sample correlation coefficients $\hat{r}_{ij}$'s are still not homoscedastic.

As a result, the entry-depandent adaptive thresholding is introduced. Here, a new larger class of sparse covariance matrices (Cai and Liu, 2011) is defined by

$$\mathcal{U}^{adp}(q, c_0(p)) = \{\mathbf{\Sigma} : \max_i \sum_{j=1}^{p} (\sigma_{ii}\sigma_{jj})^{(1-q)/2}|\sigma_{ij}|^q \leq c_0(p)\}, \tag{2.3}$$

for $0 \leq q < 1$. The parameter space $\mathcal{U}^{adp}(q, c_0(p))$ contains the uniformity class $\mathcal{U}(q, c_0(p), M)$ defined in equation (2.1), and it allows the largest variance $\max_i \sigma_{ii} \to \infty$, not requiring the variances uniformly bounded by M any more. Different from the simple thresholding, the threshold ω is not universal, but changing with the entries. It is defined as

$$\omega_{ij} = \delta\sqrt{\hat{\theta}_{ij}\frac{\log p}{n}}, \qquad i, j = 1, \ldots, p,$$

where

$$\hat{\theta}_{ij} = \frac{1}{n}\sum_{t=1}^{n}[(Y_{it} - \bar{Y}^i)(Y_{jt} - \bar{Y}^j) - \hat{\sigma}_{ij}]^2, \tag{2.4}$$

with $\bar{Y}^i$ being the sample mean of the i^{th} row of the observed data matrix $\mathbf{Y}$, and $\hat{\sigma}_{ij}$ being the corresponding sample covariance element. Note that the regularization parameter δ can be chosen by cross-validation, or just be set fixed to 2 as recommonded by Cai and Liu (2011). Then the adaptive estimator of covariance matrix is $\widehat{\mathbf{\Sigma}}_{\mathrm{adp}} = (\hat{\sigma}_{ij}^{\mathrm{adp}})_{p\times p}$ with

$$\hat{\sigma}_{ij}^{\mathrm{adp}} = \begin{cases} \hat{\sigma}_{ij} & \text{for } i = j, \\ \hat{\sigma}_{ij}\mathbb{1}_{\{|\hat{\sigma}_{ij}|>\omega_{ij}\}} & \text{for } i \neq j. \end{cases}$$

Alternatively, $\widehat{\mathbf{\Sigma}}_{\mathrm{adp}} = (\hat{\sigma}_{ij}^{\mathrm{adp}})_{p\times p}$ can be defined in an universal entry independent way as

$$\hat{\sigma}_{ij}^{\mathrm{adp}} = \begin{cases} \hat{\sigma}_{ij} & \text{for } i = j, \\ \hat{\sigma}_{ij}\mathbb{1}_{\{|\hat{\sigma}_{ij}|/\hat{\theta}_{ij}^{1/2}>\omega\}} & \text{for } i \neq j, \end{cases}$$

where $\widehat{\theta}_{ij}^{1/2}$ is the estimated standard error of the estimated covariance element $\widehat{\sigma}_{ij}$ defined in equation (2.4), and ω is the universal threshold same as the simple thresholding method (Bickel and Levina, 2008a). It is shown that $\widehat{\mathbf{\Sigma}}_{\text{adp}}$ achieves the optimal rate of convergence $c_0(p)\omega^{1-q}$ over the parameter space $\mathcal{U}^{adp}(q, c_0(p))$.

2.1.3 Generalized Thresholding Function

Other than the hard thresholding used above, there are more complex thresholding functions that can be applied, for example, soft-thresholding. Rothman et al. (2009) proposed a generalized thresholding function $s_\omega : \mathbb{R} \to \mathbb{R}$, for any ω and $z \in \mathbb{R}$, satisfying the following three conditions:

1. $|s_\omega(z)| \leq |z|$;
2. $s_\omega(z) = 0$, for $|z| \leq \omega$;
3. $|s_\omega(z) - z| \leq \omega$.

Condition 1 establishes the shrinkage. Although the bias increases, the variances are reduced and the estimator is more stable. Condition 2 enforces the thresholding, and Condition 3 limits the amount of shrinkage to no more than ω. Note that the parameter ω in the last two conditions can be different. Most of the commonly used thresholding procedures satisfy these conditions, such as the hard thresholding as we discussed above, and the soft thresholding proposed by Donoho and Johnstone (1994) which is defined as

$$s_\omega^{\text{soft}}(z) = z\mathbb{1}_{\{|z|>\omega\}}. \tag{2.5}$$

Moreover, the Smoothly Clipped Absolute Deviation (SCAD) penalty proposed by Fan and Li (2001) (see Chapter 2.1.4) and Adaptive LASSO proposed by Zou (2006) are both special cases of the generalized thresholding function $s_\omega(z)$. It is shown that, uniformly on the class $\mathcal{U}(q, c_0(p), M)$, for $\omega = C\sqrt{\log p/n}$ where C is sufficiently large,

$$\|s_\omega(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}_0\| = O_p(c_0(p)\omega^{1-q}).$$

For the generalized thresholding, the corresponding covariance estimator is defined as

$$\widehat{\mathbf{\Sigma}}_{\text{gen}} = (\widehat{\sigma}_{ij}^{gen})_{p\times p}, \qquad \widehat{\sigma}_{ij} = \begin{cases} \widehat{\sigma}_{ij}, & \text{for } i = j, \\ s_\omega(\widehat{\sigma}_{ij}), & \text{for } i \neq j. \end{cases} \tag{2.6}$$

The disadvantage of thresholding is that the estimator is not positive definite for certain.

2.1.4 Penalized Likelihood

Penalized likelihood is another powerful method for exploring sparsity. Dempster (1972) recognised the inverse covariance matrix as the canonical parameter of a multivariate normal distribution, through whih the parsimony can be identified through modified Cholesky decomposition. The nonredundant entries of this matrix are the regression coefficients of one variable based on its predecessors, so that the task of modelling a covariance matrix can be reduced to that of modelling regression models (Wu and Pourahmadi, 2003), where penalized likelihood function can be used to shrink the off-diagonal elements of the matrix.

Antoniadis and Fan (2001) proposed a penalty function (2.6) with similar properties to the generalized thresholding function in Chapter 2.1.3. Considering a linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where parameter $\boldsymbol{\beta}$ is a $d \times 1$ vector, $\mathbf{y}$ and $\boldsymbol{\epsilon}$ are $n \times 1$ vectors and the design matrix $\mathbf{X}$ is a $n \times d$ matrix. $\boldsymbol{\beta}$ can be estimated by solving the penalized least squares minimization problem

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|),$$

where $p_\lambda(\cdot)$ is a penalty function with parameter λ. Naturally, if L_0 penalty is used, $p_\lambda(|\beta|) = \lambda \mathbb{1}_{\{\beta \neq 0\}}$, it will lead to the Best-Subset selection. It finds the subset of size k that gives the smallest residual sum of squares $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ for $k = 0, 1, \ldots d$. The L_1 penalty $p_\lambda(|\beta|) = \lambda|\beta|$ yields to the soft thresholding rule as stated in equation (2.5), or Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) in the general least squares and likelihood settings. Moreover, the L_2 penalty $p_\lambda(|\beta|) = \lambda|\beta|^2$ gives a ridge regression and the L_q penalty $p_\lambda(|\beta|) = \lambda|\beta|^q$ results in a bridge regression. And if we take $p_\lambda(|\beta|) = \lambda^2 - (|\beta| - \lambda)^2 \mathbb{1}_{\{|\beta| < \lambda\}}$, it will lead to the hard thresholding rule stated in equation (2.2).

The penalty functions we discussed above do not simultaneously satisfy the three properties for a good penalty function: unbiasedness, sparsity and continuity. Therefore, Fan and Li (2001) proposed a Smoothly Clipped Absolute Deviation (SCAD) penalty,

defined by

$$p'_\lambda(\beta) = \lambda\{\mathbb{1}_{\{\beta\le\lambda\}} + \frac{(a\lambda-\beta)_+}{(a-1)\lambda}\mathbb{1}_{\{\beta>\lambda\}}\},$$

for some $a > 2$ and positive β. Here $(a)_+ = \max(0, a)$. The corresponding piecewise linear thresholding function is then

$$s_\lambda(z) = \begin{cases} \text{sgn}(z)(|z|-\lambda)_+ & \text{for } |z| \le 2\lambda, \\ [(a-1)\ z - \text{sgn}(z)a\lambda]/(a-2) & \text{for } 2\lambda \le |z| \le a\lambda, \\ z & \text{for } |z| > a\lambda. \end{cases}$$

SCAD can be seen as a combination of the soft and hard thresholding penalty functions, with improvements in the properties of these two penalty functions, selecting significant variables without creating excessive biases. The amount of shrinkage decreases when the magnitude of z rises. Note that $a = 3.7$ is suggested by Fan and Li (2001).

2.2 Estimation of Covariance Matrix with Bandable Structure

Apart from assuming the sparse structure of the true covariance matrix, other popular assumptions on matrix sturcture are widely used as well, for example the bandable structure, where the elements of the matrix decay while moving away from the diagonal. Introduced by Bickel and Levina (2008b), we consider the following class of positive definite symmetric well-conditioned matrices

$$\mathcal{U}^{\text{band}}(\alpha, M, C) = \{\mathbf{\Sigma} : \max_j \sum_i \{|\sigma_{ij}| : |i-j| > k\} \le Ck^{-\alpha} \text{ for all } k > 0, \\ \text{and } 0 < 1/M \le \lambda_{\min}(\mathbf{\Sigma}) \le \lambda_{\max}(\mathbf{\Sigma}) \le M < \infty\}, \quad (2.7)$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ represent the maximal and minimal eigenvalues of the matrix. Under this assumption, regularizing sample covariance by banding, tapering and block thresholding will provide good estimators.

2.2.1 Banding

Bickel and Levina (2008b) proposed to estimate the covariance matrix by banding the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\text{sam}}$ as

$$\widehat{\mathbf{\Sigma}}_{\text{band}}(k) = (\widehat{\sigma}_{ij}\mathbb{1}_{\{|i-j|\leq k\}})_{p\times p}$$

for some $0 \leq k < p$. It can also be written as the Schur products of the sample covariance matrix and a banding matrix $\mathbf{B}_k$, defined as

$$\mathbf{B}_k = (\mathbb{1}_{\{|i-j|<k\}})_{1\leq i,j\leq p}.$$

The estimator $\widehat{\mathbf{\Sigma}}_{\text{band}}(k) = \widehat{\mathbf{\Sigma}}_{\text{sam}} \circ \mathbf{B}_k$, where $\circ$ represents the Schur coordinate-wise matrix multiplication.

This regularization works ideally when the population covariance matrix element $\sigma_{ij} = 0$ for $|i-j| > k$ after arranging the indexes i's of the observed data $y_{it}, t = 1, \ldots, n$ under some special way, for example finite inhomogenous moving average process $y_{it} = \sum_{m=1}^{k} a_{t,t-m}\epsilon_m$ with independent and identically distributed mean zero ϵ_m.

They proved that the estimator $\widehat{\mathbf{\Sigma}}_{\text{band}}(k)$ is consistent under the operator norm if $n^{-1}\log p \to 0$, uniformly over $\mathcal{U}^{\text{band}}$. The rate of convergence is $(n^{-1}\log p)^{\alpha/2(\alpha+1)}$ with parameter $k \asymp (n/\log p)^{1/2(\alpha+1)}$. Due to the lack of information on the decay rate α in real case studies, K-folded cross-validation is recommended to better choose the parameter k in banding method.

Banding method discards the off-diagonal entries of the sample covariance matrix that are k step away from the main diagonal, setting those values to zero, while keeping the sub-diagonal or super-diagonal elements near the main diagonal unchanged. It can be generalized to the 'tapering' method, where the elements gradually decay while moving away from the main diagonal of the matrix.

2.2.2 Tapering

Cai et al. (2010) suggested to estimate the covariance matrix by tapering the maximum likelihood estimator $\widehat{\mathbf{\Sigma}}_{\text{MLE}}$. For large n, $\widehat{\mathbf{\Sigma}}_{\text{MLE}}$ is very close to the sample covariance matrix. So, I will simply use $\widehat{\sigma}_{ij}$ here for the entries of $\widehat{\mathbf{\Sigma}}_{\text{MLE}}$.

The tapering estimator is defined as the Schur products of the sample covariance matrix and tapering matrix $\mathbf{T}_k = (w_{ij})_{p\times p}$ as

$$\widehat{\mathbf{\Sigma}}_{\text{taper}}(k) = \widehat{\mathbf{\Sigma}}_{\text{MLE}} \circ \mathbf{T}_k = (\widehat{\sigma}_{ij} \cdot w_{ij})_{p\times p},$$

for some even integer k with $1 \le k \le p$, and the weights

$$w_{ij} = \begin{cases} 1 & \text{for } |i-j| \le k/2, \\ 2 - \frac{|i-j|}{k/2} & \text{for } k/2 < |i-j| < k, \\ 0 & \text{for } |i-j| \ge k. \end{cases}$$

For simplicity, the weights can also be written as

$$w_{ij} = (k/2)^{-1}\{(k - |i-j|)_+ - (k/2 - |i-j|)_+\}.$$

Over the parameter space $\mathcal{U}^{\text{band}}(\alpha, M, C)$ defined in equation (2.7), this estimator $\widehat{\mathbf{\Sigma}}_{\text{taper}}(k)$ was proven to obtain the optimal rate of convergence $n^{-2\alpha/(2\alpha+1)} + (\log p)/n$ under operator norm with parameter $k \asymp n^{1/(2\alpha+1)}$.

Compared with the rate of convergence of banding estimator $\widehat{\mathbf{\Sigma}}_{\text{band}}(k)$, which was shown to be sub-optimal, Cai et al. (2010) proved the minimax rate of convergence for $\widehat{\mathbf{\Sigma}}_{\text{taper}}(k)$ is indeed optimal.

The estimator $\widehat{\mathbf{\Sigma}}_{\text{taper}}(k)$ can also be written as the sum of many small block matrices along the main diagonal, and the size of the small blocks depends on the decay rate α.

2.2.3 Block Thresholding

As banding and tapering estimators are critically depending on the decay rate α, which is unknown in practice, Cai and Yuan (2012) introduced an adaptive estimator $\widehat{\mathbf{\Sigma}}_{\text{block}}$, independent of α and obtaining the optimal rate of convergence as we discussed in last section.

Under the same class setting as equation (2.7), $\widehat{\mathbf{\Sigma}}_{\text{block}}$ is a data-driven block thresholding method constructed in two steps. First, the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\text{sam}}$ is divided into small blocks with different size. By choosing the size $k_0 = \lfloor \log p \rfloor$ of the diagonal blocks, we gradually increase the size of blocks while moving away from

the diagnal. The floor function $\lfloor a \rfloor$ represents the greatest integer less than or equal to a. Then estimating the entries in each block by thresholding simutaneously such that:

- For the blocks located in the diagonal: keep their original values;
- For the large blocks whose dimension is larger than $n/\log n$, 'kill' them by setting these blocks as 0;
- For other blocks, threshold them by a certain thresholding rule, for example, hard thresholding, soft thresholding or adaptive Lasso.

This entire data-driven estimator $\widehat{\mathbf{\Sigma}}_{\text{block}}$ benefits from no requirement to choose α, and Cai and Yuan (2012) showed that it simultaneously attains the optimal rate of convergence $n^{-2\alpha/(2\alpha+1)} + (\log p)/n$ for estimating bandable covariance matrices over the full range of the parameter spaces $\mathcal{U}^{\text{band}}(\alpha, M, C)$.

Block thresholding has a wide range of applications, such as financial stock market prices. Traditional thresholding, as we discussed in Chapter 2.1, requires the assumption of sparsity, which the stock market along with a lot types of other empirical data cannot satisfy. In financial portfolio management, simple sparsity is not realistic for some cases, as the stock prices for similar type of stocks will have high correlation. As a result, the reasonable assumption in practice is to assume a block structure of the population covariance matrix, which is that the data are divided into groups or blocks by their features.

2.3 Estimation of Covariance Matrix with Factor Analysis

The assumptions of the sparse or bandable stucture of population coariance matrix are widely used nowadays. However, these assumptions are not always perfectly fit for all applications. We take financial market analysis as an example. Under the same stock exchange or market, although each stock has its own unique risk, they share the same systematic risk of that particular exchange or market. This results in the high correlation between every pair of stocks which cannot be neglected. Apparently, sparsity and banding assumptions do not fit well in this case. Based on the fact that the stock prices are actually controlled by only a few common factors, factor analysis is then

introduced. The benefit of factor modelling comes from the reduction of dimensions, that the number of parameters needed for estimation is reduced significantly from $p \times (p+1)/2$ parameters to $p \times (K+1)$ for the best scenario, where K is the number of factors and $0 \leq K < p$.

Let $\mathbf{y}_t = (y_{1t}, \ldots, y_{pt})^{\mathrm{T}}$ be the observed data at time $t = 1, \ldots, n$, the factor model is defined by

$$\mathbf{y}_t = \mathbf{a}_t + \mathbf{B}\mathbf{f}_t + \mathbf{u}_t, \tag{2.8}$$

where $\mathbf{a}_t$ is a p-dimension vector of data means, $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)^{\mathrm{T}}$ is a $p \times K$ matrix of factor loadings, $\mathbf{f}_t$ is a vector of common factors in dimension K, and $\mathbf{u}_t = (u_{1t}, \ldots, u_{pt})^{\mathrm{T}}$ is a vector of idiosyncratic components, i.e. error or noise terms.

2.3.1 Factor Model with Observable Factors

Factor models have been widely used both theoretically and empirically in economics and finance, such as the famous Fama and French three factor model, Capital Asset Pricing Model (CAPM), and Arbitrage Pricing Theory (APT). In some applications, the factors can be observed, such as the Fama and French three factor model, in which the excess return, market capitalization and book-to-market ratio are the three factors considered. Without loss of generality, we assume the mean vector $\mathbf{a}_t = \mathbf{0}$ in equation (2.8), and the factors $\mathbf{f}_t$ are independent of the noises $\mathbf{u}_t$. The factor model covariance matrix estimator for $\mathbf{y}_t$ is

$$\mathbf{\Sigma}_{\mathrm{FM}} = \mathbf{B}\mathbf{\Sigma}_f\mathbf{B}^{\mathrm{T}} + \mathbf{\Sigma}_u, \tag{2.9}$$

where $\mathbf{\Sigma}_f$ and $\mathbf{\Sigma}_u$ are the covariance matrices for factors $\mathbf{f}_t$ and noises $\mathbf{u}_t$, respectively.

If we further assume the cross-sectional independence among the idiosyncratic components, which implies that the noise covariance matrix $\mathbf{\Sigma}_u = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$ for $\mathbf{u}_t$ is diagonal, we call this a strict factor model (Fan et al., 2008). Here, $\mathrm{diag}(a_1, \ldots, a_m)$ denote the diagonal matrix of dimension m whose diagonal elements equal to $a_1, \ldots, a_m$ and off-diagonal elements are all zero. After regressing on $\mathbf{y}_t$, estimation of loading matrix $\widehat{\mathbf{B}}$ and the residuals $\widehat{\mathbf{\Sigma}}_u = \mathrm{diag}(\widehat{\sigma}_1, \ldots, \widehat{\sigma}_p)$ can be obtained. $\widehat{\mathbf{B}} = (\widehat{\mathbf{b}}_1, \ldots, \widehat{\mathbf{b}}_p)^{\mathrm{T}}$ can be estimated by applying the least squares estimation as

$$\widehat{\mathbf{b}}_i = \arg\min_{\mathbf{b}_i} \frac{1}{n}\sum_{t=1}^{n}(y_{it} - \mathbf{b}_i^{\mathrm{T}}\mathbf{f}_t)^2, \quad \text{for } i = 1, \ldots, p,$$

and $\hat{\mathbf{u}}_t = \mathbf{y}_t - \hat{\mathbf{B}}\mathbf{f}_t$. Since the factors $\mathbf{f}_t$'s are observable, $\hat{\mathbf{\Sigma}}_f$ is just the sample covariance matrix for the factors. The resulting estimator for strict factor model is given by

$$\hat{\mathbf{\Sigma}}_{\text{SFM}} = \hat{\mathbf{B}}\hat{\mathbf{\Sigma}}_f\hat{\mathbf{B}}^{\text{T}} + \hat{\mathbf{\Sigma}}_u.$$

For any matrix $\mathbf{A}$, the Frobenius norm is defined as $\|\mathbf{A}\|_F = \{\text{tr}(\mathbf{A}\mathbf{A}^{\text{T}})\}^{1/2}$. Under the Frobenius norm, the dimensionality reduces the rate of convergence by the order of pK, same as that of the sample covariance matrix. Since the sample covariance matrix achieving the optimal rate as well, the factor structure does not give many advantages in estimating the population covariance matrix.

On the other hand, in the aspect of estimating the precision matrix $\mathbf{\Sigma}_0^{-1}$, it can be estimated by using the Sherman-Morrison-Woodbury formula as

$$\hat{\mathbf{\Sigma}}_{\text{SFM}}^{-1} = \hat{\mathbf{\Sigma}}_u^{-1} - \hat{\mathbf{\Sigma}}_u^{-1}\hat{\mathbf{B}}(\hat{\mathbf{\Sigma}}_f^{-1} + \hat{\mathbf{B}}^{\text{T}}\hat{\mathbf{\Sigma}}_u^{-1}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^{\text{T}}\hat{\mathbf{\Sigma}}_u^{-1}.$$

Fan et al. (2008) proved that $\hat{\mathbf{\Sigma}}_{\text{SFM}}^{-1}$ performs much better than $\hat{\mathbf{\Sigma}}_{\text{sam}}^{-1}$ when the number of factor $K = O(p)$.

Compared to the sample covariance matrix, $\hat{\mathbf{\Sigma}}_{\text{SFM}}$ gets better convergence rate for estimating the precision matrix $\mathbf{\Sigma}_0^{-1}$ but the same rate for estimation of $\mathbf{\Sigma}_0$. As a result, factor modelling performs better in portfolio allocation, but not much for risk assessment.

Strict factor model requires the chosen factors explaining all relationships between the data, which is too restricted to satisfy in practical sense. The approximate factor model estimator $\hat{\mathbf{\Sigma}}_{\text{AFM}}$ is introduced by Fan et al. (2011), loosing the strict diagonal structrue of error covariance matrix to an approximate diagonal structrue. This allows the existence of small valued off-diagonal entries in estimated idiosyncratic matrix, suggesting the allowance of the presence of cross-sectional correlation even after taking out the common factors. The estimator is defined as

$$\hat{\mathbf{\Sigma}}_{\text{AFM}} = \hat{\mathbf{B}}\hat{\mathbf{\Sigma}}_f\hat{\mathbf{B}}^{\text{T}} + \tilde{\mathbf{\Sigma}}_u.$$

Similar to strict factor model, $\hat{\mathbf{B}}$ can be estimated by least squares method. Since the factors are all observable, $\hat{\mathbf{\Sigma}}_f$ is estimated by the sample covariance matrix of $\mathbf{f}_t$, and $\hat{\mathbf{\Sigma}}_u$ is the sample covariance of $\hat{\mathbf{u}}_t = \mathbf{y}_t - \hat{\mathbf{B}}\mathbf{f}_t$. As the estimated covariance matrix for the idiosyncratic terms is no longer a diagonal matrix but a sparse matrix, adaptive

thresholding (Cai and Liu, 2011) is used to regularize $\widehat{\mathbf{\Sigma}}_u = (\widehat{\sigma}_{u,ij})_{p\times p}$ as

$$\widetilde{\mathbf{\Sigma}}_u = (\widetilde{\sigma}_{u,ij})_{p\times p}, \qquad \widetilde{\sigma}_{u,ij} = \begin{cases} \widehat{\sigma}_{u,ij} & i = j, \\ s_\omega(\widehat{\sigma}_{u,ij}) & i \neq j, \end{cases}$$

where $s_\omega(\cdot)$ is the general thresholding function and here adaptive threshoding $s_\omega(\widehat{\sigma}_{u,ij}) = \widehat{\sigma}_{u,ij}\mathbb{1}_{\{|\widehat{\sigma}_{u,ij}|>\omega_{ij}\}}$ is applied in their paper (see Chapter 2.1.2 for adaptive thresholding and Chapter 2.1.3 for generalized thresholding function).

Fan et al. (2011) proved that the estimated covariacne matrix is still invertible after thresholding even if $p > n$. It is clearly shown that p can be much larger than n when estimating the precision matrix. Eigenvalues of $\widehat{\mathbf{\Sigma}}_{\text{AFM}}$ diverge quickly while those of $\widehat{\mathbf{\Sigma}}^{-1}_{\text{AFM}}$ are uniformly bounded, suggesting a good performance of the estimated precision matrix also for approximate factor structure.

2.3.2 Factor Model with Unobservable Factors

In many practical applications, the common factors are always unobservable, i.e. latent. So estimation of the latent factors are needed. We apply the pervasive assumption that the number of factors K are bounded and eigenvalues of $p^{-1}\mathbf{B}^{\intercal}\mathbf{B}$ are uniformly bounded away from 0 and ∞ as $p \to \infty$.

We cannot observe the factors directly since they are latent. As the dimension p increases, the information about the common factors accumulates while that about the error terms does not. It helps to distinguish the factor term $\mathbf{Bf}_t$ from the idiosyncratic term $\mathbf{u}_t$. As a result, the principal component analysis related method is widely used in this field.

When the number of variables, i.e. dimension p, is large, we can use Principal Component Analysis (PCA) to analyze the factor model. We assume that $\mathbf{a}_t = \mathbf{0}$ for simplicity, and factor model in equation (2.8) is then

$$\mathbf{y}_t = \mathbf{Bf}_t + \mathbf{u}_t = (\mathbf{BH})(\mathbf{H}^{-1}\mathbf{f}_t) + \mathbf{u}_t,$$

where $\mathbf{H}$ is a $K \times K$ non-singular matrix. The solution for the pair $(\mathbf{H}, \mathbf{f}_t)$ is not unique, caused by the problem of 'identifiability' due to the unknown $\mathbf{B}$ and $\mathbf{f}_t$. To solve the ambiguity of the solution, we further assume that the covariance of $\mathbf{f}_t$ is a K dimensional identity matrix such that $\mathbf{\Sigma}_f = \mathbf{I}_K$, and the columns of $\mathbf{B}$ are orthogonal.

Similar to equation (2.9), the covariance matrix is defined as

$$\mathbf{\Sigma}_{\text{PC}} = \mathbf{B}\mathbf{\Sigma}_f\mathbf{B}^{\text{T}} + \mathbf{\Sigma}_u = \mathbf{B}\mathbf{B}^{\text{T}} + \mathbf{\Sigma}_u,$$

where $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_K)$. The K orthogonal columns $\mathbf{b}_j$, $j = 1, \ldots, K$ are in a descending order according to the the norms $\|\mathbf{b}_j\|$. For simplicity, we assume a strictly descending order of $\|\mathbf{b}_j\|$. It is easy to get the j-th eigenvector of $\mathbf{B}$ is the normalized vector $\mathbf{b}_j/\|\mathbf{b}_j\|$. The corresponding j-th largest eigenvalue is $\|\mathbf{b}_j\|^2$ for $j < K$, and 0 for the remaining $p - K$ eigenvalues since

$$(\mathbf{B}\mathbf{B}^{\text{T}})\mathbf{b}_j/\|\mathbf{b}_j\| = (\sum_{i=1}^{K} \mathbf{b}_j\mathbf{b}_j^{\text{T}})\mathbf{b}_j/\|\mathbf{b}_j\| = \|\mathbf{b}_j\|^2\mathbf{b}_j/\|\mathbf{b}_j\|.$$

If the j-th factor satisfies the pervasive assumption, the eigenvalue $\|\mathbf{b}_j\|^2 = \sum_{i=1}^{p} b_{ij}^2$ is of order p, and j-th factor influences a non-negligible fraction of the data $\mathbf{Y}$ among the p dimensions.

It is well known from linear algebra that the first K eigenvalues of $\mathbf{B}\mathbf{B}^{\text{T}}$ are the same as those of $\mathbf{B}^{\text{T}}\mathbf{B}$. If $\mathbf{\Sigma}_u = \mathbf{0}$ and $\mathbf{\Sigma}_{\text{PC}} = \mathbf{B}\mathbf{B}^{\text{T}}$, the j-th eigenvlue of $\mathbf{\Sigma}_{\text{PC}} = \mathbf{B}\mathbf{B}^{\text{T}}$ is

$$\lambda_j = \begin{cases} \|\mathbf{b}_j\|^2 & \text{for } j \leq K, \\ 0 & \text{for } j > K, \end{cases} \tag{2.10}$$

and the corresponding eigenvector is

$$\mathbf{p}_j = \mathbf{b}_j/\|\mathbf{b}_j\|, \tag{2.11}$$

making the covariance matrix be just the simple version of the spectral decompsition

$$\mathbf{\Sigma}_{\text{PC}} = \sum_{j=1}^{K} \lambda_j\mathbf{p}_j\mathbf{p}_j^{\text{T}}.$$

If we allow a more general case of $\mathbf{\Sigma}_u$, the eigenvalues and eigenvectors in equations (2.10) and (2.11) only hold approximately. By the Wely's theorem and the $\sin(\theta)$ theorem of Davis and Kahan (1970), we have that $\left\|\mathbf{p}_j - \mathbf{b}_j/\|\mathbf{b}_j\|\right\| = O(p^{-1}\|\mathbf{\Sigma}_u\|)$ and

$\left|\lambda_j - \|\mathbf{b}_j\|^2\right| \leq \|\mathbf{\Sigma}_u\|$ for $j \leq K$, and $|\lambda_j| \leq \|\mathbf{\Sigma}_u\|$ for $j > K$. As a result,

$$\mathbf{\Sigma}_0 \approx \sum_{j=1}^{K} \lambda_j \mathbf{p}_j \mathbf{p}_j^{\mathrm{T}} + \mathbf{\Sigma}_u.$$

Under factor model assumptions, the first K eigenvalues of $\mathbf{\Sigma}_0$ are very spiked, while the rest are either bounded or growing slowly. So the latent factors and the loadings can be approximated using the eigen-decomposition of $\mathbf{\Sigma}_0$. As a result, high-dimensional factor model can be estimated by the principal component analysis.

To take advantage of the factor structure estimation and the sparseness exploitation in the error covariance matrix, Fan et al. (2013) proposed the Principal Orthogonal complEment Thresholding (POET) method for the approximate factor structure with sparsity, allowing the presence of some cross-sectional correlations.

POET uses the principal component analysis on the sample covariance matrix, and then thresholds the idiosyncratic matrix. Consider a factor model like equation (2.8) with $\mathbf{a}_t = \mathbf{0}$. The $p \times p$ covariance matrix of $\mathbf{y}_t$ is given by $\mathbf{\Sigma}_0 = \mathbf{B}\mathbf{\Sigma}_f\mathbf{B}^{\mathrm{T}} + \mathbf{\Sigma}_u$. POET assumes an approximate factor model that the first K eigenvalues of $\mathbf{\Sigma}_0$ are spiked and grow at a rate $O(p)$, and $\mathbf{\Sigma}_u$ is approximately sparse.

Now let λ_i and $\mathbf{p}_i$ be the descending ordered eigenvalues and corresponding eigenvectors of the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\text{sam}}$, and K be the number of diverging eigenvalues of $\mathbf{\Sigma}_0$. Then, the spectral decomposition of $\widehat{\mathbf{\Sigma}}_{\text{sam}}$ becomes

$$\widehat{\mathbf{\Sigma}}_{\text{sam}} = \sum_{i=1}^{K} \lambda_i \mathbf{p}_i \mathbf{p}_i^{\mathrm{T}} + \widehat{\mathbf{\Sigma}}_u,$$

where $\widehat{\mathbf{\Sigma}}_u = \sum_{i=K+1}^{p} \widehat{\lambda}_i \widehat{\mathbf{p}}_i \widehat{\mathbf{p}}_i^{\mathrm{T}} = (\widehat{\sigma}_{u,ij})_{p\times p}$ is the principal orthogonal complement. Due to the conditional sparsity assumption, they threshold on $\widehat{\mathbf{\Sigma}}_u$ as

$$\widetilde{\mathbf{\Sigma}}_u = (\widetilde{\sigma}_{u,ij})_{p\times p}, \quad \widetilde{\sigma}_{u,ij} = \begin{cases} \widehat{\sigma}_{u,ii}, & i = j; \\ s_\omega(\widehat{\sigma}_{u,ij}), & i \neq j, \end{cases}$$

where $s_\omega(.)$ is a generalized thresholding function (see Chapter 2.1.3), and the adaptive thresholding $s_\omega(z_{ij}) = z_{ij}\mathbb{1}_{\{|\widehat{z}_{ij}| \geq \omega_{ij}\}}$ is suggested here. The POET estimator is then

$$\widehat{\mathbf{\Sigma}}_{\text{POET}} = \sum_{i=1}^{K} \lambda_i \mathbf{p}\mathbf{p}^{\mathrm{T}} + \widetilde{\mathbf{\Sigma}}_u.$$

When $K = 0$, this estimator is the same as the simple adaptive thresholding (Cai and Liu, 2011) or more general thresholding cases depending on the choice of thresholding function (Rothman et al., 2009).

The estimator has an equivalent representation using a constrained least squares method with results $\widehat{\mathbf{B}} = (\widehat{\mathbf{b}}_1, \ldots, \widehat{\mathbf{b}}_p)$ and $\widehat{\mathbf{f}}_t$. Let $\widehat{u}_{it} = y_{it} - \widehat{\mathbf{b}}_i^{\mathrm{T}}\widehat{\mathbf{f}}_t$, $\widehat{\sigma}_{u,ij} = n^{-1}\sum_{t=1}^{n}\widehat{u}_{it}\widehat{u}_{jt}$, and denote $\widehat{\theta}_{ij} = n^{-1}\sum_{t=1}^{n}(\widehat{u}_{it}\widehat{u}_{jt} - \widehat{\sigma}_{u,ij})^2$. The adaptive threshold parameter for $s_\omega(\cdot)$ applied here is

$$\omega_{ij} = C\sqrt{\widehat{\theta}_{ij}} \cdot \left(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}}\right),$$

where $C > 0$ is a sufficiently large constant.

Fan et al. (2013) proved that the estimated idiosyncratic matrix $\widetilde{\mathbf{\Sigma}}_u$ and POET estimator $\widehat{\mathbf{\Sigma}}_{\mathrm{POET}}$ are consistent estimators for $\mathbf{\Sigma}_u$ and $\mathbf{\Sigma}_0$ respectively. With the increase of dimension p, more information can be provided. It is easier to distinguish the factors with the noise, making this unknown factor case equivalent to the known factor one.

The advantage of this low rank with sparse matrix estimation is that POET is simple, optimization-free and it uses the data only through the sample covariance matrix.

For most empirical applications, the number of factors K is unknown. The multi-fold cross-validation method can be used to estimate K. After determining the estimated parameter $\widehat{K}$, the POET method above then can be applied

$$\widehat{\mathbf{\Sigma}}_{\mathrm{POET}} = \sum_{i=1}^{\widehat{K}} \lambda_i \mathbf{p}\mathbf{p}^{\mathrm{T}} + \widetilde{\mathbf{\Sigma}}_u.$$

2.4 Estimation of Covariance Matrix by Shrinkage

The estimation methods discussed in previous sections all require a beforehand knowledge of covariance matrix structure. Although these methods perform well under their own assumed structure, the estimators would still go far away from the true matrix if this prior information is inaccurate. In real world cases, we usually cannot know the true covariance matrix structure in advance, leading to the difficulty of choosing the correct matrix structure or estimation method.

Researchers begin to find estimation methods that do not require assumptions on the structure of the true covariance matrix. Simple sample covariance matrix estimation does not assume any special structure, but the drawback is well-known that it is ill-conditioned when the dimension p is large relative to the sample size n. Here, we introduce four cutting edge researches developing covariance estimation by shrinking the sample eigenvalues.

Stein (1956) first used this idea to estimate the mean vector. This shrinkage method is substituting the original ill-conditioned estimator with a convex combination of itself and a target matrix. It balances between the estimation error coming from the ill-conditioned variance matrix and the specification error associated with the target matrix.

2.4.1 Linear Shrinkage

There is one type of covariance matrix estimation combining linearly two or even more existing estimators. For example, Ledoit and Wolf (2004) proposed a well-conditioned estimator $\widehat{\mathbf{\Sigma}}_{\text{lin}}$, which is the asymptotically optimal convex linear combination of the sample covariance matrix with the identity matrix under quadratic loss function.

Here, we define the condition number as the ratio of the maximal and minimal singular values of the estimator. If the condition number is not much larger than one, we say the matrix is well-conditioned, which means its inverse can be computed with good accuracy. If the condition number is very large, the matrix is said to be ill-conditioned. Practically, such a matrix is almost singular, and the computation of its inverse or the solution of a linear system of equations is more likely to cause large numerical errors. A non-invertible matrix has condition number equal to infinity. As a result, well-condition is a very important property that a good estimator should obtain.

Sample covariance matrix works well only if $p \ll n$. For the estimation of large covariance matrices, p is large and it is difficult to find enough samples with size n to make p/n negligible. While $p/n < 1$ but not negligible, the estimator may be invertible, but will be ill-conditioned. Sample covariance matrix is worse-conditioned than the true covariance matrix. Researchers showed that sample eigenvalues are more dispersed around their grand mean than the true ones, and the excess dispersion is equal to the error of sample covariance matrix. Excess dispersion implies that the largest sample eigenvalues are biased upwards and the smallest ones downwards. To make an estimator

well-conditioned, a direct method is to force it to be well-conditioned in structure by diagonality or factor model for example. But the absence of prior information would lead the structure to be mis-specified in general.

To shrink sample covariance towards to the identity matrix, Ledoit and Wolf (2004) introduced a linear combination of sample covariance matrix and the identity matrix as,

$$\widehat{\mathbf{\Sigma}}_{\text{lin}} = w \cdot \text{tr}(\widehat{\mathbf{\Sigma}})/p \cdot \mathbf{I}_p + (1-w) \cdot \widehat{\mathbf{\Sigma}}_{\text{sam}}, \quad \text{where}$$
$$w = \min\Big(1, \frac{\frac{1}{n^2}\sum_{i=1}^{n} \|\mathbf{y}_i\mathbf{y}_i^{\text{T}} - \widehat{\mathbf{\Sigma}}\|_F^2}{\|\widehat{\mathbf{\Sigma}} - \text{tr}(\widehat{\mathbf{\Sigma}})/p\|_F^2}\Big).$$

Here, $\mathbf{I}_p$ is the identity matrix with dimension p, and w is the shrinkage intensity measuring the amount of shrinkage of $\widehat{\mathbf{\Sigma}}_{\text{sam}}$ towards $\mathbf{I}_p$. This weighted average of sample covariance and identity matrix is equivalent to linearly shrink the sample eigenvalues to the grand mean while retaining the sample eigenvectors.

In the general asymptotics framework, where $p/n \to c > 0$, the optimal shrinkage intensity w can be estimated consistently. This estimator $\widehat{\mathbf{\Sigma}}_{\text{lin}}$ benefits from its free of distribution and computational advantage. Simulation results show that $\widehat{\mathbf{\Sigma}}_{\text{lin}}$ outperforms the sample covariance matrix and does well in finite sample version. More importantly, $\widehat{\mathbf{\Sigma}}_{\text{lin}}$ is better-conditioned, and always be invertible even when $p > n$.

2.4.2 Nonlinear Shrinkage on Eigenvalues

Different from their previous paper we discussed in last section applying the one-size-fits-all approach (Ledoit and Wolf, 2004), Ledoit and Wolf (2012) proposed a new rotation equivariant estimator focusing on the individualized shrinkage intensity to every sample eigenvalue. A covariance matrix estimator is rotation equivariant if and only if it has the same eigenvectors as the sample covariance matrix. As such an estimator, it can only differentiate itself by its eigenvalues.

To do this, the eigen-decomposition is applied in the sample covariance matrix by

$$\widehat{\mathbf{\Sigma}}_{\text{sam}} = \mathbf{P}\mathbf{D}\mathbf{P}^{\text{T}},$$

where $\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_p)$ and $\mathbf{D} = \text{diag}(\lambda_1, \ldots, \lambda_p)$ are the $p \times p$ matrix of eigenvectors and p-dimension digonal matrix of all eigenvalues of $\widehat{\mathbf{\Sigma}}_{\text{sam}}$, respectively.

Under large-dimensional asymptotics $p/n \to c \in (0, 1)$ as $n \to \infty$, define H_n as the empirical distribution function of the population covariance matrix and F_n as that of the sample covariance matrix. Assume H_n converges to some limit H. Using the Stieltjes transformation $m_{F_n}(z), \forall z \in \mathbb{C}^+$, the Marcenko-Pastur equation $m_F(z)$ of function F_n and the estimation of $\widetilde{m}_F(\lambda) \equiv \lim_{z \to \lambda} m_F(z)$, they proposed the estimator through the minimization problem under Frobenius loss

$$\min_{\mathbf{A}} \left\| \mathbf{PAP}^{\mathrm{T}} - \mathbf{\Sigma}_0 \right\|_F,$$

where $\mathbf{A} = \text{diag}(a_1, \ldots, a_p)$ is the p-dimension diagonal matrix. Ledoit and Péché (2011) showed the solution for the minimization problem $a_i, i = 1, \ldots, p$ can be approximated by

$$\widetilde{a}_i = \frac{\lambda_i}{|1 - c - c\lambda_i \widetilde{m}_F(\lambda_i)|^2},$$

where $\widetilde{a}_i$ only depends on the limiting distribution of sample eigenvalues. Hence, they proposed an estimator $\widehat{\mathbf{\Sigma}}_{\text{nonlin}}$ through nonlinearly shrinkage of the sample eigenvalues:

$$\widehat{\mathbf{\Sigma}}_{\text{nonlin}} = \mathbf{P} \cdot \text{diag}\left(\widetilde{a}_i\right)_{i=1,\ldots,p} \cdot \mathbf{P}^{\mathrm{T}}.$$

Simulation study shows a significant improvement over the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\text{sam}}$ and the linear shrinkage estimator $\widehat{\mathbf{\Sigma}}_{\text{lin}}$ (Ledoit and Wolf, 2004) when the sample size n is very large compared to the dimension p.

2.4.3 Condition Number Regularized Estimator

Based on the fact that orthogonal matrices are never ill-conditioned, Abadir et al. (2014) proposed a new rotation equivariant estimator by applying orthogonal decomposition of the sample covariance matrix under high-dimensional setting $p/n \to 0$.

The condition number of any orthogonal matrix is always equal to 1, so the ill-condition of the sample covariance matrix only comes from the eigenvalues part of $\widehat{\mathbf{\Sigma}}_{\text{sam}}$. As a result, our focus should be on the improvement of the sample eigenvalues

$\lambda_i, i = 1, \ldots, p$. The estimated diagonal matrix of eigenvalues is

$$\mathbf{D} = \mathbf{P}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_{\mathrm{sam}}\mathbf{P} = \mathrm{diag}(\mathrm{var}(\mathbf{p}_1^{\mathrm{T}}\mathbf{Y}), \ldots, \mathrm{var}(\mathbf{p}_p^{\mathrm{T}}\mathbf{Y})). \tag{2.12}$$

Now, instead of using the whole dataset $\mathbf{Y}$ to estimate $\mathbf{P}$, only m ($m < n$) observations are used to approximately orthogonalize the rest of the $n - m$ observations, which are used to re-estimate $\mathbf{D}$. Assume $m \to \infty$ and $n - m \to \infty$ as $n \to \infty$, and $\mathbf{y}_i$'s are i.i.d. distributed. The reason of spliting the whole sample set is that $\mathbf{D}$ cannot reuse the data that has already been used for calculating $\mathbf{P}$, since they worsen the estimate of $\mathbf{D}$. Denote the whole data as $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$, where $\mathbf{Y}_1$ has dimension $p \times m$ and $\mathbf{Y}_2$ has dimension $p \times (n - m)$. Here, $\widetilde{\mathbf{\Sigma}}_1$ and $\widetilde{\mathbf{\Sigma}}_2$ are the sample covariance of $\mathbf{Y}_1$ and $\mathbf{Y}_2$, respectively. Define $\widetilde{\mathbf{\Sigma}}_1 = \mathbf{P}_1\mathbf{D}_1\mathbf{P}_1^{\mathrm{T}}$, where $\mathbf{P}_1$ is the matrix of eigenvectors of $\mathbf{Y}_1$. The new estimator for eigenvalues is then defined as

$$\widehat{\mathbf{D}} = \mathrm{diag}(\mathrm{var}(\mathbf{P}_1^{\mathrm{T}}\mathbf{Y}_2)) = \mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{P}_1),$$

and the resulting Condition number Regularized Covariance matrix estimator (CRC) is then defined as

$$\widehat{\mathbf{\Sigma}}_{\mathrm{CRC}} = \mathbf{P}\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{P}_1)\mathbf{P}^{\mathrm{T}}. \tag{2.13}$$

Abadir et al. (2014) further improved the estimator by repeatly choose different subsamples ($\mathbf{Y}_1$ and $\mathbf{Y}_2$) and taking the average of the estimators by all cases. For simulation study, the split point is selected by balancing the trade-off between variance and bias through bootstraps.

Such an estimator reduces the multi-variate problem to p univariate ones, enabling also the easy-to-apply estimation of the functions on $\mathbf{\Sigma}_0$, such as the estimation of the precision matrix. $\widehat{\mathbf{\Sigma}}_{\mathrm{CRC}}$ is also always design-free and well-conditioned even when $p > n$.

2.4.4 NERCOME

Inspired by Abadir et al. (2014), Lam (2016) proposed a covariance matrix estimator that nonparametrically regularize the eigenvalues (NERCOME).

They first proved the theoretical properties of the regularized eigenvalues in Abadir et al. (2014). When the observations can be written as $\mathbf{y}_i = \mathbf{\Sigma}_0^{1/2}\mathbf{z}_i$, where $\mathbf{z}_i$'s are independently and identically distributed entries, the regularized eigenvalues in $\widehat{\mathbf{\Sigma}}_{\mathrm{CRC}}$

are asymptotically the same as the nonlinearly shrunk ones in $\widehat{\mathbf{\Sigma}}_{\text{nonlin}}$ (Ledoit and Wolf, 2012). The computer advantage for $\widehat{\mathbf{\Sigma}}_{\text{CRC}}$ cones from the only involvement of eigen-decompositions of $p \times p$ matrices, while $\widehat{\mathbf{\Sigma}}_{\text{nonlin}}$ requiring nonconvex optimizations could be computational expansive.

Lam (2016) also showed that the nonlinear shrinkage formula in Ledoit and Wolf (2012) is not correct if the data is from factor model, due to the low dimensional factors. The data splitting regularized eigenvalues are still asymtotically optimal when we consider the Frobenius loss minimization.

Relaxing from the high-dimensional setting Abadir et al. (2014) that $p/n \to 0$, Lam (2016) considered $p/n \to c > 0$. They focused on the optimization problem

$$\min_{\mathbf{D}} \|\mathbf{P}_1 \mathbf{D} \mathbf{P}_1^{\mathrm{T}} - \mathbf{\Sigma}_0\|_F,$$

and the optimal solution for $\mathbf{D} = \text{diag}(d_1, \ldots, d_p)$ is

$$d_i = \mathbf{p}_{1i}^{\mathrm{T}} \mathbf{\Sigma}_0 \mathbf{p}_{1i}.$$

Through splitting the data into two independent parts, the estimated eigenvalues are the diagonal elements of matrix $\mathbf{P}_1^{\mathrm{T}} \widetilde{\mathbf{\Sigma}}_2 \mathbf{P}_1$, same as those in Abadir et al. (2014). The estimator is similar to $\widehat{\mathbf{\Sigma}}_{\text{CRC}}$ in equation (2.13), expect that $\mathbf{P}$ is substituted by $\mathbf{P}_1$, the matrix of eigenvectors for the first m of the whole data, $\mathbf{Y}_1$, only. They proposed the estimator as

$$\widehat{\mathbf{\Sigma}}_m = \mathbf{P}_1 \text{diag}(\mathbf{P}_1^{\mathrm{T}} \widetilde{\mathbf{\Sigma}}_2 \mathbf{P}_1) \mathbf{P}_1^{\mathrm{T}}.$$

The change from $\mathbf{P}$ to $\mathbf{P}_1$ makes sense due to their optimization problem under Frobenius loss and they showed that the $\mathbf{p}_{1i}^{\mathrm{T}} \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i}$ is asytotically the same as d_i. Apparently, this estimator is not as informative as $\widehat{\mathbf{\Sigma}}_{\text{CRC}}$, due to the change of $\mathbf{P}_1$. As a result, they announced the final estimator as the averaging over all $\widehat{\mathbf{\Sigma}}_m$'s calculated by different choices of $\mathbf{Y}_1$ and $\mathbf{Y}_2$. Referring to the assumption that each vector $\mathbf{y}_i$ in $\mathbf{Y}$ is independent of each other and identically distributed, the data $\mathbf{Y}$ can be permuted so that there can be multiple choices of data matrix $\mathbf{Y}_1^{(j)}$ and $\mathbf{Y}_2^{(j)}$. The corresponding covariance estimator for j^{th} permutation is $\widehat{\mathbf{\Sigma}}_m^{(j)}$ and the final NERCOME estimator is

$$\widehat{\mathbf{\Sigma}}_{\text{NERCOME}} = \frac{1}{M} \sum_{j=1}^{M} \widehat{\mathbf{\Sigma}}_m^{(j)}.$$

They showed theoretically and empirically better performance of $\widehat{\mathbf{\Sigma}}_{\text{NERCOME}}$, compared to $\widehat{\mathbf{\Sigma}}_{\text{CRC}}$. The regularization of eigenvalues gives NERCOME asymptotically optimal nonlinear shrinkage with respect to Frobenius norm. $\widehat{\mathbf{\Sigma}}_{\text{NERCOME}}$ is positive definite almost surely as long as the population covariance $\mathbf{\Sigma}_0$ is, even in a high-dimensional setting ($p > n$).

2.4.5 NOVELIST

The linear combination of two covriance estimators can be considered as an other kind of nonlinear eigenvalue shrinkage. Huang and Fryzlewicz (2015) proposed a NOVEL Integration of the Sample and Thresholded covariance estimators (NOVELIST), performing shrinkage of the sample covariance / correlation towards its thresholded version.

Denote $\widehat{\mathbf{R}}_{\text{sam}}$ as the sample correlation matrix, and $s_\lambda(\cdot)$ as the generalized thresholding function (Rothman et al., 2009) with threshold λ (see Chapter 2.1.3). The correlation version of NOVELIST is defined as

$$\widehat{\mathbf{R}}_{\text{NOVELIST}} = (1-\delta)\cdot\widehat{\mathbf{R}}_{\text{sam}} + \delta\cdot s_\lambda(\widehat{\mathbf{R}}_{\text{sam}}),$$

where δ is the weight parameter for the thresholded estimator. The corresponding covariance version of NOVELIST is defined as $\widehat{\mathbf{\Sigma}}_{\text{NOVELIST}} = \mathbf{D}^{1/2}\widehat{\mathbf{R}}_{\text{NOVELIST}}\mathbf{D}^{1/2}$, with $\mathbf{D}$ as the diagnal matrix the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\text{sam}}$.

The parameters (λ, δ) do not have a theoretical optimal solution, so that they are obtained by cross-validation or just a fixed set of parameter choice based on different scenarios in data analysis. Thanks to the flexible control of the degree of shrinkage and thresholding, NOVELIST, as a simple, good all-round covariance, correlation and precision matrix estimator, offer competitive performance across other covariance matrix estimation methods.

2.5 Estimation of Covariance Matrix in High Frequency Setting

With the rapid development in technology, people are not satisfied with the limited amount of data any more. The demand of big data analysis becomes higher and higher.

For example, in stock market, there are thousands of transactions for a stock every day, not even include the tons of bid and ask data. Why would investors only use the daily closing price, the low frequency data, for analysis? The dramatical increase in data will benefit the research through bringing lots of extra information, but it will also cause some problems. In this section, the problems caused by high-frequency data and possible solutions are discussed.

2.5.1 Integrated Variance

First, we only consider the simplest case, where the dimension equals to one. When the input data's frequency increases, the common variance is no longer useful, as we are more interested in the total variation over a certain period, i.e. the integrated variance. The most commonly used estimator is the Realized Variance (RV). Let X_t be the price process for stock return in logarithm, following an Ito process,

$$dX_t = \mu_t dt + \sigma_t dW_t, \quad t \in [0, 1],$$

where μ_t and σ_t are the drift and volatility terms which can follow some random processes, and B_t is a standard Brownian motion. Under the continuous setting, the quadratic variation, i.e. the true variance of our interest, is given by

$$[X] = \int_0^1 \sigma_t^2 dt.$$

To estimate this variance, the commonly used method is called the realized variance or the quadratic covariation, defined as

$$[X] = \lim_{n\to\infty} \sum_{i=1}^{n} (X_{t_i} - X_{t_{i-1}})^2, \tag{2.14}$$

for any sequence of partitions $0 = t_0 < t_1 < \cdots < t_n = 1$ with $\sup_i(t_i - t_{i-1}) \to 0$ for $n \to \infty$. Here X_{t_i} represents the i^{th} transaction price during the periond $[0, 1]$, and t_i is the time when this i^{th} transaction happens.

Theoretically, this method would work, like the sample variance for low-frequency data. However, in real case analysis, problems occur. Since the existence of noises, the prices which we can observe are not the true underlying prices, but the ones contanminated by some noise like the bid-ask spread. We call this noise the market

microstructure noise. Let Y_{t_i} be the log-price we observe at time t_i, it is equal to the sum of true underlying log-price X_{t_i} and the microstructure noise ϵ_{t_i}:

$$Y_{t_i} = X_{t_i} + \epsilon_{t_i},$$

where ϵ is independent of X.

As a result, we can only get the realized volatility using all the data in Y, $[Y]^{\text{all}}$, but not $[X]$ in equation (2.14) that we need. The contamination of microstructure noise will grow with the increase of data frequency, due to the fact that the change in true returns gets smaller while the noise remains at the same magnitude.

A natural solution is to choose a proper frequency interval to have a trade-off between the extra information and the unwanted extra noise. So we will have a sparse set of data based on Y. The realized volatility based on this sparse data is denoted as $[Y]^{\text{sparse}}$. We can further improve the method by choosing the interval according to the minimization of the Mean Squared Error (MSE). Although the sparse method would keep the error in a relatively low level, but it discards a lot of data!

A reasonable solution is to get different sparse subsamples of data, calculated $[Y]^{\text{sparse}}$ separately and then averaging all the subsamples' realized volatilities. Denote this as $[Y]^{\text{avg}}$. It will benefit from the low level of noise due to the sparse dataset, and gain as much information as possible because of the use of different subsamples in the same time.

All the estimations try to decrease the variance of the noise, but they are still biased. Zhang et al. (2005) proposed a Two-Scale Realized Volatility (TSRV) method that combining $[Y]^{\text{all}}$ and $[Y]^{\text{avg}}$ together. TSRV is defined as

$$\widehat{\langle X \rangle} = [Y]^{\text{avg}} - \frac{\bar{n}}{n}[Y]^{\text{all}},$$

where n is the sample size of whole Y, and $\bar{n} = (n - K + 1)/K$ for K being the number of subgrids.

This method benefits from the rich sources of tick-by-tick data and corrects for the adverse effects of microstructure noise on volatiltiy estimation to a great extent.

Besides TSRV, the recent literature on realized volatility also includes works by Barndorff-Nielsen and Shephard (2002), Meddahi (2002), Andersen et al. (2003), and Hansen and Lunde (2006).

2.5.2 Integrated Covariance Matrix

If we increase the dimesion to two or more, the Integrated CoVariance (ICV) matrix is of interest. Let $\mathbf{X}_t = (X_t^{(1)}, \cdots, X_t^{(p)})^{\mathrm{T}}$ be a p-dimensional log-price diffusion process modeled by

$$d\mathbf{X}_t = \boldsymbol{\mu}_t dt + \boldsymbol{\sigma}_t d\mathbf{W}_t, \ t \in [0,1],$$

where $\boldsymbol{\mu}_t$ is the drift, $\boldsymbol{\sigma}_t$ is a $p \times p$ matrix of instantaneous covolatility process, and $\mathbf{W}_t$ is a p-dimensional standard Brownian motion. We want to estimate the integrated covariance matrix, defined by

$$\boldsymbol{\Sigma}_0 = \int_0^1 \boldsymbol{\sigma}_t \boldsymbol{\sigma}_t^{\mathrm{T}} dt.$$

A so-called Realized CoVariance (RCV) matrix is widely used to estimate this ICV, when the observed data are synchronous at high frequency. RCV is defined as

$$\boldsymbol{\Sigma}_{\mathrm{RCV}} = \sum_{i=1}^{n} \Delta\mathbf{X}_\ell \Delta\mathbf{X}_i^{\mathrm{T}}, \quad \text{where } \Delta\mathbf{X}_i = \mathbf{X}_{t_i} - \mathbf{X}_{t_{i-1}}.$$

In real cases, for example the financial applications with the contamination of noises, RCV usually is not the perfect estimator for the desired true covariance. Actually more problems will come out, such as the nonsynchronized trading time and the Epps effect (Epps, 1979). The transactions take place discretely at different times for the two assets, as well as the noise obtain. As a result, the correlation estimates between assets tend to decrease for high frequencies trading. For two log-price series $X_t^{(1)}$ and $X_t^{(2)}$ following Ito processes,

$$\begin{aligned} dX_t^{(1)} &= \mu_t^{(1)} dt + \sigma_t^{(1)} dW_t^{(1)}, \\ dX_t^{(2)} &= \mu_t^{(2)} dt + \sigma_t^{(2)} dW_t^{(2)}, \end{aligned}$$

where $W_t^{(1)}$ and $W_t^{(1)}$ are both standard Brownian motions with correlation ρ_t at time t. The integrated covariation during time $[0,1]$ will be

$$\left\langle X^{(1)}, X^{(2)} \right\rangle = \int_0^1 \sigma_t^{(1)} \sigma_t^{(2)} d\left\langle W^{(1)}, W^{(2)} \right\rangle_t.$$

To solve the non-synchronous problem, a refresh time method (Barndorff-Nielsen et al., 2011) is applied. From some starting point (the previous refresh time v_{i-1}), the current refresh time v_i is that when both assets are traded at least once. The refresh

times for two assets are defined as the last transaction time before or at this refresh time, t_i and s_i for $X^{(1)}$ and $X^{(2)}$ respectively. Similar to TSRV, Zhang (2011) proposed the Two-Scales realized CoVariance (TSCV) estimator using observed data $Y^{(1)}$ and $Y^{(2)}$ as

$$\widehat{\langle X^{(1)}, X^{(2)} \rangle} = c \cdot ([Y^{(1)}, Y^{(2)}]^K - \frac{n_K}{n_J}[Y^{(1)}, Y^{(2)}]^J), \tag{2.15}$$

where $[Y^{(1)}, Y^{(2)}]^K = \frac{1}{K}\sum_{i=K}^{n}(Y^{(1)}_{t_i} - Y^{(1)}_{t_{i-K}})(Y^{(2)}_{s_i} - Y^{(2)}_{s_{i-K}})$ is the averaged lag K previous-tick realized covariance, $c = 1 + o_p(n^{-1/6})$ is a constant, $K = O(n^{2/3})$ and $1 \leq J \ll K$. In the classical two scales setting, J is usually set as 1.

For positively associated assets $X^{(1)}$ and $X^{(2)}$, a negative bias is introduced to the estimator due to the non-synchronous trading. By applying TSCV, the two scale estimation could eliminate the bias due to asynchronicity and microstructure noise, achieving better performance in high-frequency setting without loss of information.

Apart from TSCV, Multi-Scale Realized Volatility Matrix (MSRVM) by Tao et al. (2013), the Kernel Realized Volatility Matrix (KRVM) by Barndorff-Nielsen et al. (2011) and the Pre-averaging Realized Volatility Matrix (PRVM) by Christensen et al. (2010) are also giving good performance for estimating the covariance matrix under the high-frequency setting.

Chapter 3

Integrating Regularized Covariance Matrix Estimators

3.1 Introduction

Estimation of a covariance matrix or its inverse, precision matrix, is an important and sometimes inevitable task in data analysis. Thanks to the richness of data we can obtain nowadays, more often than not the data are high-dimensional in natrue, creating problems for traditional estimators, such as the sample covariance matrix. A particularly serious problem is that the eigenvalues of the sample covariance matrix are more extreme than their population counterpart, as well documented in Bai and Yin (1993) and subsequent random matrix theory researches (see Bai and Silverstein (2010) for example). Moreover, when the dimension p is larger than the sample size n, the sample covariance matrix is not invertible.

As reviewed in Chapter 2, two major branches of regularization methods are developed for covariance matrix estimation in view of the problems above. One branch assumes that the population covariance matrix $\mathbf{\Sigma}_0$ or the corresponding precision matrix $\mathbf{\Sigma}_0^{-1}$ has special structures. Sparseness of $\mathbf{\Sigma}_0$ is one of the most commonly exploit structure in applications (Bickel and Levina, 2008a; Cai and Yuan, 2012; Cai and Zhou, 2012; Lam and Fan, 2009; Rothman et al., 2009). Sparseness of $\mathbf{\Sigma}_0^{-1}$ is closely connected to graphical modelling (Friedman et al., 2008, Meinshausen and Bühlmann, 2006). Banded structure of $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_0^{-1}$ (i.e., only limited number of off-diagonals and the main diagonal are non-zero) are treated in Bickel and Levina

(2008b). If the data follows a factor model, then $\mathbf{\Sigma}_0$ can have a factor structure (Fan et al., 2008, 2011). Fan et al. (2013) combines factor structure estimation as well as sparseness exploitation in the residual covariance matrix.

Another branch concerns with regularizing, or "shrinking", the eigenvalues of the sample covariance matrix under the high-dimensional setting $p/n \to c > 0$. Ledoit and Wolf (2004) proposed the linear shrinkage estimator, which is a weighted average of the identity and the sample covariance matrix, shrinking the eigenvalues towards a grand mean. This is generalized to shrinkage towards a specified matrix "target" other than the identity in Schäfer and Strimmer (2005). Won et al. (2013) regularized on the condition number, winsorizing the extreme eigenvalues of the sample covariance matrix at certain constants. Ledoit and Wolf (2012, 2013) proposed a rotation-equivariant estimator with nonlinear shrinkage of eigenvalues, while Lam (2016), using a data-splitting idea from Abadir et al. (2014), proved that such nonlinear shrinkage can be achieved through data-splitting with theoretical justification of the split location. Recently, Huang and Fryzlewicz (2015) proposed the NOVELIST, a weighted average of the sample covariance matrix and a thresholded estimator, and obtained good practical results. This can also be considered as a form of nonlinear shrinkage estimator, although there are no theoretical justifications in the exact form of shrinkage.

While both branches of researches provide good estimators under different scenarios, there are no methods that can "take advantage" of what each branch of estimators can offer. For instance, if the sparse assumption on $\mathbf{\Sigma}_0$ is correct, then a thresholded estimator $\mathbf{T}$ is good. Yet, $\mathbf{T}$ can be far from $\mathbf{\Sigma}_0$ if the sparse assumption turns out to be only roughly true or not true at all. Even when the sparse assumption is correct, finite sample performance of $\mathbf{T}$ can still be not as good as a shrinkage estimator. With this in mind, a desirable estimator is one that can be automatically similar to a structured estimator $\mathbf{T}$ when it is indeed close to $\mathbf{\Sigma}_0$ in a certain sense, and be similar to a shrinkage estimator automatically when $\mathbf{T}$ is not performing well. Moreover, when $\mathbf{\Sigma}_0$ is approximately banded and approximately sparse for instance, we hope that a regularized covariance matrix estimator can take advantages from both banded and thresholded estimators at the same time, and be able to "switch" to one particular estimator if that is close enough to $\mathbf{\Sigma}_0$ in a certain sense.

To bridge the gap described above, we propose an estimator that combines two or even three regularized covariance estimators through a weighted average. This is not dissimilar with the NOVELIST in Huang and Fryzlewicz (2015), except that we

are combining a rotation-equivariant estimator, instead of just the sample covariance matrix, with other regularized estimators. The freedom of the diagonals in the rotation-equivariant estimator, together with the weight on each regularized estimator, allows for a flexible final estimator. One major contribution in this chapter is that, by minimizing the Frobenius loss, we derive explicit formulae for the weights which can be estimated consistently, or even almost surely, through a data splitting scheme which is similar to the one in Lam (2016). These weights indeed allow our estimator to have the desirable property that it approaches a particular regularized estimator $\mathbf{T}$ if it is indeed close to $\mathbf{\Sigma}_0$ in a certain sense, and approaches the rotation-equivariant estimator if other regularized estimators are not performing well. See Chapter 3.2 and Theorem 3.2 and 3.5 for more details. Extensive simulations also reveal that our estimator can indeed gather the advantages from a regularized estimator and perform well, even when the regularized estimator itself does not. As another contribution, we show that our estimator is asymptotically efficient when compared to an ideal estimator constructed with the knowledge of $\mathbf{\Sigma}_0$. Such efficiency proof, similar to Lam (2016), also provides possible theoretical split locations, and give insights into choosing a practical one for the data. Finally, extension to even more regularized estimators is not difficult using the mathematics behind the proof of Theorem 3.1 and 3.4, although we do not pursue in this direction in this chapter.

The rest of this chapter is organized as follows. Chapter 3.2 defines the concept of our estimator, and shows the form of an ideal estimator through Frobenius loss minimization. A bona fide estimator is introduced in Chapter 3.2.2 with data splitting, and its theoretical properties presented in Chapter 3.2.3. Chapter 3.3 extends everything to two regularized covariance matrices. Chapter 3.4.1 introduces an averaged estimator which has better performance and is more stable, while detailing the practical procedures in the choice of split location for the data and other practical concerns. Chapter 3.5 presents the extensive simulation results with a real data analysis as well. Finally Chapter 3.6 includes all the proof of the theorems in this chapter.

3.2 Proposed Estimator with a Single Regularized Matrix

We consider a covariance matrix estimator of the form

$$\boldsymbol{\Sigma}(\delta, \mathbf{D}) = (1-\delta)\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}} + \delta\mathbf{T}, \tag{3.1}$$

where $\mathbf{T}$ is a regularized covariance matrix using a chosen regularization method, and $\mathbf{P}$ is an orthogonal matrix. The above estimator is weighted between the regularized estimator $\mathbf{T}$ and the matrix $\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}}$. Lam (2016) used a rotation-equivariant estimator $\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}}$ as the basis of a covariance matrix estimator, where $\mathbf{P}$ contains all the eigenvectors of a sample covariance matrix constructed from the available data. Similar to Abadir et al. (2014), a sample splitting scheme is used in Lam (2016) to find the diagonal matrix $\mathbf{D}$.

Our estimator closely resembles the NOVEL Integration of the Sample and Thresholded covariance estimators (NOVELIST) proposed in Huang and Fryzlewicz (2015), in the sense that our estimator also aims to integrate with a regularized estimator, for instance, a thresholded covariance matrix estimator $\mathbf{T}$, using a linear weighting scheme. As such, if $\mathbf{T}$ is a fixed "target" matrix which does not depend on data, it also resembles the linear shrinkage estimator proposed in Ledoit and Wolf (2004), where a linear combination of the sample covariance matrix and a fixed target matrix is considered.

The major difference, however, is that the sample covariance matrix in both papers is replaced by a rotation-equivariant estimator $\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}}$, where $\mathbf{D}$ is diagonal matrix to be determined. Hence, we are integrating two regularized covariance matrices, instead of just regularizing the sample covariance matrix through linear weighting with another regularized estimator. The major motivation in doing so is that while the rotation-equivariant estimator can help achieve nonlinear shrinkage of eigenvalues as in Lam (2016) without assuming a specific structure of the true covariance matrix $\boldsymbol{\Sigma}_0$, it may lose accuracy against a regularized estimator which assumes a specific structure for $\boldsymbol{\Sigma}_0$, if such an assumption does ultimately hold. In this sense, we hope to achieve a better estimator through $\boldsymbol{\Sigma}(\delta, \mathbf{D})$ when we are not certain if a particular assumption on $\boldsymbol{\Sigma}_0$ holds. Ideally, the weight δ should be close to 1 if $\mathbf{T}$ is regularized correctly assuming a specific structure on $\boldsymbol{\Sigma}_0$, and close to 0 if $\mathbf{T}$ is too far away from the true one

due to a wrong assumption on $\mathbf{\Sigma}_0$. We show that this is indeed the case asymptotically almost surely using our proposed estimator (3.5) in Chapter 3.2.2 below.

3.2.1 Frobenius Loss Minimization

We propose to estimate δ and $\mathbf{D}$ through minimizing the Frobenius loss. Ledoit and Wolf (2004) also used Frobenius loss minimization for their linear shrinkage estimator. Hence, we consider

$$\min_{\delta,\mathbf{D}} \|(1-\delta)\mathbf{PDP}^{\mathrm{T}} + \delta\mathbf{T} - \mathbf{\Sigma}_0\|_F^2, \tag{3.2}$$

where $\|\mathbf{A}\|_F^2 = \mathrm{tr}(\mathbf{AA}^{\mathrm{T}})$, with $\mathrm{tr}(\mathbf{A})$ being the trace of a matrix $\mathbf{A}$. We present the solution to the minimization problem above in the following theorem.

Theorem 3.1 *Suppose $\delta \neq 1$ and $\mathbf{T}$ is not of the form $\mathbf{PDP}^{\mathrm{T}}$ for some diagonal matrix $\mathbf{D}$. Define $\widehat{\mathbf{\Sigma}}_{\mathbf{T}} = \mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{TP})\mathbf{P}^{\mathrm{T}}$, where $\mathrm{diag}(\mathbf{A})$ represents a diagonal matrix with diagonal entries as in the matrix $\mathbf{A}$. Then the solution to the minimization problem (3.2) is*

$$\mathbf{D} = \frac{1}{1-\delta}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{P}) - \frac{\delta}{1-\delta}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{TP}), \text{ with } \delta = \frac{\mathrm{tr}[(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})\mathbf{\Sigma}_0]}{\mathrm{tr}(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})^2}.$$

The requirement that $\mathbf{T}$ cannot be of the form $\mathbf{PDP}^{\mathrm{T}}$ is a regularity condition. If this condition is not satisfied, then $\mathrm{tr}(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})^2 = 0$, and the original problem actually reduces to the one considered in Lam (2016) or Ledoit and Wolf (2012). The same thing happens if $\mathbf{\Sigma}_0 = \sigma^2\mathbf{I}_p$, since then $\mathrm{tr}((\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})\mathbf{\Sigma}_0) = 0$, and so $\delta = 0$. This condition is only required by theoretical proofs but not have a bearing in practice (see Chapter 3.4.2 for details). Substituting the forms of $\mathbf{D}$ and δ into $\mathbf{\Sigma}(\delta, \mathbf{D})$, the corresponding covariance matrix estimator depends on $\mathbf{P}, \mathbf{T}$ and $\mathbf{\Sigma}_0$, and is given by

$$\mathbf{\Sigma}(\mathbf{P}, \mathbf{T}, \mathbf{\Sigma}_0) = \mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}} + \frac{\mathrm{tr}[(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})\mathbf{\Sigma}_0]}{\mathrm{tr}(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})^2}(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}}). \tag{3.3}$$

The first part of the estimator coincides with the "ideal" nonlinear shrinkage estimator in Ledoit and Wolf (2012) and Lam (2016) for efficiency comparisons purpose. The second part is a weighted version of $\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}}$, which is itself congruent to a hollow matrix, in the sense that $\mathbf{P}(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})\mathbf{P}^{\mathrm{T}}$ has zero diagonal. The weight $\mathrm{tr}[(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})\mathbf{\Sigma}_0]/\mathrm{tr}(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})^2$ can actually go negative or above 1. Looking at its formula, it is a generalized angle

between $\mathbf{\Sigma}_0$ and $\mathbf{T} - \widehat{\mathbf{\Sigma}}_\mathbf{T}$, calculated using the inner product $\text{tr}(\mathbf{AB})$ for symmetric real square matrices $\mathbf{A}$ and $\mathbf{B}$. When the inner product $\text{tr}[(\mathbf{T} - \widehat{\mathbf{\Sigma}}_\mathbf{T})\mathbf{\Sigma}_0] = 0$, it means that $\mathbf{\Sigma}_0$ and $\mathbf{T} - \widehat{\mathbf{\Sigma}}_\mathbf{T}$ are "orthogonal" to each other in a generalized sense, and hence $\mathbf{T} - \widehat{\mathbf{\Sigma}}_\mathbf{T}$ is not important and is therefore weighted 0, leaving only the ideal nonlinear shrinkage estimator $\mathbf{P}\text{diag}(\mathbf{P}^\text{T}\mathbf{\Sigma}_0\mathbf{P})\mathbf{P}^\text{T}$. Obviously we cannot use this estimator in practice since it depends on $\mathbf{\Sigma}_0$ itself.

Remark 3.1 *Theorem 3.1 does not allow $\delta = 1$. However, as $\delta \to 1$, using $tr[(\mathbf{T} - \widehat{\mathbf{\Sigma}}_\mathbf{T})\widehat{\mathbf{\Sigma}}_\mathbf{T}] = 0$, we can easily deduce from the form of δ in the result of Theorem 3.1 that the estimator $\mathbf{\Sigma}(\mathbf{P}, \mathbf{T}, \mathbf{\Sigma}_0)$ approaches the form $\mathbf{T} + \mathbf{P}\,diag(\mathbf{P}^\text{T}(\mathbf{\Sigma}_0 - \mathbf{T})\mathbf{P})\mathbf{P}^\text{T}$. This makes sense, since even if $\mathbf{T}$ is a good estimator in itself so that we assume $\delta \to 1$ in equation (3.1), in terms of minimizing the Frobenius loss, $\mathbf{T} + \mathbf{P}\,diag(\mathbf{P}^\text{T}(\mathbf{\Sigma}_0 - \mathbf{T})\mathbf{P})\mathbf{P}^\text{T}$ is always better than $\mathbf{T}$ alone by easy calculation.*

3.2.2 Proposed Estimator with Data Splitting

Since $\mathbf{\Sigma}_0$ is unknown, we need to propose a practical estimator which can be constructed from the data $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, where each observed vector $\mathbf{y}_i$ is of dimension p. Hence $\mathbf{\Sigma}_0$ is a $p \times p$ matrix. Hereafter we assume that the $\mathbf{y}_i$'s are independent of each other, each with mean $\mathbf{0}$ and covariance $\mathbf{\Sigma}_0$.

We use the sample splitting idea used in Abadir et al. (2014) and Lam (2016), and split the data into two independent portions $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$, where $\mathbf{Y}_1$ is of size $p \times m$ and $\mathbf{Y}_2$ of size $p \times (n - m)$. Construct two sample covariance matrices, and perform an eigen-decomposition on the first one:

$$\widetilde{\mathbf{\Sigma}}_1 = m^{-1}\mathbf{Y}_1\mathbf{Y}_1^\text{T} = \mathbf{P}_1\mathbf{D}_1\mathbf{P}_1^\text{T}, \quad \widetilde{\mathbf{\Sigma}}_2 = (n - m)^{-1}\mathbf{Y}_2\mathbf{Y}_2^\text{T}.$$

We then consider the Frobenius loss minimization problem

$$\min_{\delta, \mathbf{D}} \|(1 - \delta)\mathbf{P}_1\mathbf{D}\mathbf{P}_1^\text{T} + \delta\mathbf{T} - \mathbf{\Sigma}_0\|_F^2, \tag{3.4}$$

where $\mathbf{T}$ here is constructed from $\mathbf{Y}_1$ only instead of the whole data matrix $\mathbf{Y}$. This results in the estimator $\mathbf{\Sigma}(\mathbf{P}_1, \mathbf{T}, \mathbf{\Sigma}_0)$ similar to that in equation (3.3). We then propose our estimator to be $\mathbf{\Sigma}(\mathbf{P}_1, \mathbf{T}, \widetilde{\mathbf{\Sigma}}_2)$, that is,

$$\mathbf{\Sigma}(\mathbf{P}_1, \mathbf{T}, \widetilde{\mathbf{\Sigma}}_2) = \mathbf{P}_1\text{diag}(\mathbf{P}_1^\text{T}\widetilde{\mathbf{\Sigma}}_2\mathbf{P}_1)\mathbf{P}_1^\text{T} + \frac{\text{tr}[(\mathbf{T} - \widehat{\mathbf{\Sigma}}_\mathbf{T})\widetilde{\mathbf{\Sigma}}_2]}{\text{tr}(\mathbf{T} - \widehat{\mathbf{\Sigma}}_\mathbf{T})^2}(\mathbf{T} - \widehat{\mathbf{\Sigma}}_\mathbf{T}). \tag{3.5}$$

Note here $\widehat{\mathbf{\Sigma}}_{\mathbf{T}} = \mathbf{P}_1 \text{diag}(\mathbf{P}_1^{\mathrm{T}}\mathbf{T}\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}}$. The first part of this estimator is exactly the NERCOME introduced in Lam (2016) for nonlinear shrinkage of the eigenvalues of the sample covariance matrix. The salient feature that gives this estimator its nice properties (to be introduced in later theorems) is that both $\mathbf{P}_1$ and $\mathbf{T}$ are independent of $\widetilde{\mathbf{\Sigma}}_2$ by our construction. For the theoretical value of the split m and a practical one, see Chapter 3.2.3 and Chapter 3.4 respectively.

3.2.3 Theoretical Results with Single Regularized Estimator

To present the asymptotically properties of our estimator $\mathbf{\Sigma}(\mathbf{P}_1, \mathbf{T}, \widetilde{\mathbf{\Sigma}}_2)$, we list our assumptions below.

(A1) The observed vectors are independent of each other, and each $\mathbf{y}_i$ can be written as $\mathbf{y}_i = \mathbf{\Sigma}_0^{1/2}\mathbf{z}_i$ for $i = 1, \ldots, n$, where each $\mathbf{z}_i$ is a $p \times 1$ vector of independent and identically distributed random variables z_{ij}. Each z_{ij} has mean 0 and variance 1, with $E|z_{ij}|^k \leq B < \infty$ for some constant B and $2 < k \leq 20$.

(A2) The population covariance matrix is non-random and of size $p \times p$. Furthermore, $\|\mathbf{\Sigma}_0\| = O(1)$.

(A3) Let $\tau_{n,1} \geq \ldots \geq \tau_{n,p}$ be the p eigenvalues of $\mathbf{\Sigma}_0$, with corresponding eigenvectors $\mathbf{v}_{n,1}, \ldots, \mathbf{v}_{n,p}$. Define $H_n(\tau) = p^{-1}\sum_{i=1}^{p} \mathbb{1}_{\{\tau_{n,i} \leq \tau\}}$ the empirical distribution function (e.d.f.) of the population eigenvalues . We assume $H_n(\tau)$ converges to some non-random limit H at every point of continuity of H.

(A4) The support of H defined above is the union of a finite number of compact intervals bounded away from zero and infinity. Also, there exists a compact interval in $(0, +\infty)$ that contains the support of H_n for each n.

For data coming from a factor model, with

$$\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \boldsymbol{\epsilon}_i, \ i = 1, \ldots, n,$$

we still assume that the $\mathbf{y}_i$'s are independent of each other. Furthermore, we assuming the following.

(F1) The series $\{\boldsymbol{\epsilon}_i\}$ has $\boldsymbol{\epsilon}_i = \mathbf{\Sigma}_\epsilon^{1/2}\boldsymbol{\xi}_i$, where $\boldsymbol{\xi}_i$ is a $p \times 1$ vector of independent and identically distributed random variables ξ_{ij}. Each ξ_{ij} has mean 0 and unit

variance, and $E|\xi_{ij}|^k \leq B < \infty$ for some constant B and $k \leq 20$. The factor series $\{\mathbf{x}_t\}$ has a constant dimension r, and $\mathbf{x}_t = \mathbf{\Sigma}_x^{1/2}\mathbf{x}_t^*$ where $\mathbf{x}_t^*$ is a $r \times 1$ vector of independent and identically distributed random variables x_{tj}^*. Also, $E|x_{tj}^*|^k \leq B < \infty$ for some constant B and $2 < k \leq 20$.

(F2) The covariance matrix $\mathbf{\Sigma}_x = \text{var}(\mathbf{x}_i)$ is such that $\|\mathbf{\Sigma}_x\| = O(1)$. The covariance matrix $\mathbf{\Sigma}_\epsilon = \text{var}(\boldsymbol{\epsilon}_i)$ also has $\|\mathbf{\Sigma}_\epsilon\| = O(1)$. Both matrices are non-random. The factor loading matrix $\mathbf{A}$ is such that $\|\mathbf{A}\|_F^2 = O(p)$.

Assumptions (A1) and (A2) here coincide with assumptions (A1)' and (A2)', while (A3) and (A4) coincide with (A3) and (A4) in Lam (2016). They are needed for proving the asymptotic efficiency for our estimator $\mathbf{\Sigma}(\mathbf{P}_1, \mathbf{T}, \widetilde{\mathbf{\Sigma}}_2)$. For data from a factor model, (F1) is assumption (F1)', while (F2) coincides with (F2) in Lam (2016). We first present some asymptotic properties of the estimated weight

$$\widehat{\delta} = \frac{\text{tr}[(\mathbf{T} - \widehat{\mathbf{\Sigma}}_\mathbf{T})\widetilde{\mathbf{\Sigma}}_2]}{\text{tr}(\mathbf{T} - \widehat{\mathbf{\Sigma}}_\mathbf{T})^2}, \tag{3.6}$$

which uses $\mathbf{Y}_1$ for the construction of $\mathbf{T}$, and substitutes $\mathbf{\Sigma}_0$ in the result of Theorem 3.1 by $\widetilde{\mathbf{\Sigma}}_2$.

Theorem 3.2 *Let Assumptions (A1)-(A4) be satisfied, and the split location $m = m(n)$ satisfies the constraint $\sum_{n\geq 1} p(n-m)^{-5} < \infty$ while $p = p(n)$ satisfies $p/n \to c > 0$. If $p^{-1}\text{tr}(\mathbf{T} - \widehat{\mathbf{\Sigma}}_\mathbf{T})^2 \nrightarrow 0$ and is finite in probability/almost surely, and δ is defined in Theorem 3.1 with $\mathbf{T}$ constructed using $\mathbf{Y}_1$ and $\widehat{\mathbf{\Sigma}}_\mathbf{T} = \mathbf{P}_1\text{diag}(\mathbf{P}_1^\text{T}\mathbf{T}\mathbf{P}_1)\mathbf{P}_1^\text{T}$, then we have $\widehat{\delta} - \delta \to 0$ in probability/almost surely.*

Furthermore, supposing $\mathbf{\Sigma}_0 \neq \sigma^2\mathbf{I}_p$, if we have $\|\mathbf{T} - \mathbf{\Sigma}_0\| \to 0$ in probability/almost surely, then $\widehat{\delta} \to 1$ and $\|\mathbf{\Sigma}(\mathbf{P}_1, \mathbf{T}, \widetilde{\mathbf{\Sigma}}_2) - \mathbf{\Sigma}_0\| \to 0$ in probability/almost surely.

For data from a factor model, let Assumptions (F1) and (F2) be satisfied. Assume the split location $m = m(n)$ satisfies the constraint $\sum_{n\geq 1} p(n-m)^{-5} < \infty$ while $p = p(n)$ satisfies $p/n \to c > 0$. If $p^{-2}\text{tr}(\mathbf{T} - \widehat{\mathbf{\Sigma}}_\mathbf{T})^2 \nrightarrow 0$ and is finite in probability/almost surely, then we have $\widehat{\delta} - \delta \to 0$ in probability/almost surely.

Furthermore, if $p^{-1}\|\mathbf{T} - \mathbf{\Sigma}_0\| \to 0$ in probability/almost surely, then $\widehat{\delta} \to 1$ and $p^{-1}\|\mathbf{\Sigma}(\mathbf{P}_1, \mathbf{T}, \widetilde{\mathbf{\Sigma}}_2) - \mathbf{\Sigma}_0\| \to 0$ in probability/almost surely.

Our estimated weight approaches the true weight using data $\mathbf{Y}_1$ if $\mathbf{T}$ is significantly different from the form $\mathbf{P}_1\mathbf{D}\mathbf{P}_1^\text{T}$, so that $p^{-1}\text{tr}(\mathbf{T} - \widehat{\mathbf{\Sigma}}_\mathbf{T})^2$ is not going to 0. If $\mathbf{T}$

approaches $\mathbf{\Sigma}_0$ in the spectral norm, then $\mathbf{T}$ is a good enough estimator, and the theoretical result matches our intuition, that $\widehat{\delta}$ should go to 1, and favor $\mathbf{T}$ completely. The resulting estimator then approaches $\mathbf{\Sigma}_0$ in the spectral norm as well.

For data from a factor model, $\mathbf{\Sigma}_0$ is spiked with a few eigenvalues of order p from the assumptions. Then the theorem says that if $\mathbf{T}$ can estimate those spiked eigenvalues accurately enough such that $p^{-1}\|\mathbf{T}-\mathbf{\Sigma}_0\|$ goes to 0, we still favor $\mathbf{T}$ over a rotation-equivariant estimator completely.

These results are very useful in practice. For example, if $\mathbf{\Sigma}_0$ is banded, then a banded estimator $\mathbf{T}$ will have $\|\mathbf{T}-\mathbf{\Sigma}_0\|$ go to 0 in probability at a certain rate (see Bickel and Levina (2008b) for details). It means that if we are using $\mathbf{\Sigma}(\mathbf{P}_1, \mathbf{T}, \widetilde{\mathbf{\Sigma}}_2)$ as our estimator, then it asymptotically approaches $\mathbf{T}$ itself since $\widehat{\delta}$ approaches 1, which is desirable. For data from a factor model, POET from Fan et al. (2013) can serve as $\mathbf{T}$ and our results above can be applied as well.

To present efficiency properties of our estimator, we define the ideal estimator for our purpose of comparison as

$$\mathbf{\Sigma}_{\text{Ideal}} = \mathbf{\Sigma}(\mathbf{P}, \mathbf{T}, \mathbf{\Sigma}_0),$$

which is the same as the one in equation (3.3) except that $\mathbf{T}$ here is defined as the one constructed using $\mathbf{Y}_1$. The efficiency loss of an estimator $\widehat{\mathbf{\Sigma}}$ can then be defined as

$$EL(\mathbf{\Sigma}_0, \widehat{\mathbf{\Sigma}}) = 1 - \frac{\|\mathbf{\Sigma}_{\text{Ideal}} - \mathbf{\Sigma}_0\|_F^2}{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0\|_F^2}. \tag{3.7}$$

When $\widehat{\mathbf{\Sigma}}$ has a larger Frobenius loss then the ideal estimator, the efficiency loss is positive, and is negative vice versa.

Theorem 3.3 *Let Assumptions (A1)-(A4) be satisfied, and* $\mathbf{\Sigma}_0 \neq \sigma^2\mathbf{I}_p$. *Assume the split location* $m = m(n)$ *satisfies the constraints* $m/n \to 1, n-m \to \infty$ *and* $\sum_{n\geq 1} p(n-m)^{-5} < \infty$ *while* $p = p(n)$ *satisfies* $p/n \to c > 0$. *If* $p^{-1}\text{tr}(\mathbf{T} - \widehat{\mathbf{\Sigma}}_{\mathbf{T}})^2 \nrightarrow 0$ *and is finite in probability/almost surely, then* $EL(\mathbf{\Sigma}_0, \mathbf{\Sigma}(\mathbf{P}_1, \mathbf{T}, \widetilde{\mathbf{\Sigma}}_2)) \to 0$ *in probability/almost surely.*

We do not attempt to prove asymptotic efficiency when the data is from a factor model in this chapter. We illustrate the corresponding results in our simulations instead, when a regularized covariance matrix estimator like POET from Fan et al. (2013) is

used as $\mathbf{T}$ in equation (3.5). Note that $\mathbf{\Sigma}_0 \neq \sigma^2\mathbf{I}_p$ in the theorem. As discussed earlier, the problem actually reduces to the one in Lam (2016) when $\mathbf{\Sigma}_0 = \sigma^2\mathbf{I}_p$, and our estimator reduces to NERCOME too in Lam (2016), which still works well empirically (see the simulation results in Lam (2016) for more details).

3.3 Extension to Two Regularized Matrices

When we are truly uncertain what assumptions to make on $\mathbf{\Sigma}_0$, it is natural for us to try more than one regularization methods. In this chapter we introduce two regularized estimators to be included in a linear combination with a rotation-equivariant estimator. This way, we do not need to determine what regularization methods are more appropriate in advance if, like the results in Theorem 3.2, the weights are going to 0 or 1 matching our intuition when one particular regularization is more appropriate than the other.

Given two regularized estimators $\mathbf{T}_1$ and $\mathbf{T}_2$ and a fixed orthogonal matrix $\mathbf{P}$ (we use the eigenmatrix from the eigen-decomposition of the sample covariance matrix), we consider

$$\min_{\delta_1,\delta_2,\mathbf{D}} \|(1-\delta_1-\delta_2)\mathbf{PDP}^{\mathrm{T}} + \delta_1\mathbf{T}_1 + \delta_2\mathbf{T}_2 - \mathbf{\Sigma}_0\|_F^2. \tag{3.8}$$

Theorem 3.4 *Suppose $\delta_1+\delta_2 \neq 1$, and $\mathbf{T}_1$, $\mathbf{T}_2$ and $\mathbf{T}_1-\mathbf{T}_2$ are not of the form $\mathbf{PDP}^{\mathrm{T}}$ for some diagonal matrix $\mathbf{D}$. Then defining $\widehat{\mathbf{\Sigma}}_{\mathbf{T}_i} = \mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{T}_i\mathbf{P})\mathbf{P}^{\mathrm{T}}$, the solution to the minimization problem (3.8) is*

$$\delta_1 = \frac{\mathrm{tr}[(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]\mathrm{tr}(\mathbf{T}_2-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})^2 - \mathrm{tr}[(\mathbf{T}_2-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})]\mathrm{tr}[(\mathbf{T}_2-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})\mathbf{\Sigma}_0]}{\mathrm{tr}(\mathbf{T}_2-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})^2\mathrm{tr}(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})^2 - \mathrm{tr}^2[(\mathbf{T}_2-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})]},$$

$$\delta_2 = \frac{\mathrm{tr}[(\mathbf{T}_2-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})\mathbf{\Sigma}_0]\mathrm{tr}(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})^2 - \mathrm{tr}[(\mathbf{T}_2-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})]\mathrm{tr}[(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]}{\mathrm{tr}(\mathbf{T}_2-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})^2\mathrm{tr}(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})^2 - \mathrm{tr}^2[(\mathbf{T}_2-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})]},$$

$$\mathbf{D} = \frac{1}{1-\delta_1-\delta_2}\left(\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{P}) - \delta_1\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{T}_1\mathbf{P}) - \delta_2\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{T}_2\mathbf{P})\right).$$

The regularity conditions are sound, since the violation of one or more of these reduces the problem to either (3.2) or even the one considered in Lam (2016). Same thing happens when $\mathbf{\Sigma}_0 = \sigma^2\mathbf{I}_p$, since then $\delta_1 = \delta_2 = 0$. And, the denominator in δ_1 and δ_2 is 0 when any one of the regularity conditions are violated. The resulting estimator is

given by

$$\boldsymbol{\Sigma}(\mathbf{P}, \mathbf{T}, \boldsymbol{\Sigma}_0) = \mathbf{P}\text{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}} + \delta_1(\mathbf{T}_1 - \widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_1}) + \delta_2(\mathbf{T}_2 - \widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_2}). \tag{3.9}$$

3.3.1 Proposed Estimator and Theoretical Results

We use the same data splitting idea as in Chapter 3.2.2, and consider the minimization problem

$$\min_{\delta_1,\delta_2,\mathbf{D}} \|(1-\delta_1-\delta_2)\mathbf{P}_1\mathbf{D}\mathbf{P}_1^{\mathrm{T}} + \delta_1\mathbf{T}_1 + \delta_2\mathbf{T}_2 - \boldsymbol{\Sigma}_0\|_F^2,$$

where $\mathbf{T}_1$ and $\mathbf{T}_2$ are both constructed from the data matrix $\mathbf{Y}_1$. Substituting $\boldsymbol{\Sigma}_0$ by $\widetilde{\boldsymbol{\Sigma}}_2$ in the resulting solution, we have our estimator

$$\boldsymbol{\Sigma}(\mathbf{P}_1, \mathbf{T}, \widetilde{\boldsymbol{\Sigma}}_2) = \mathbf{P}_1\text{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}} + \widehat{\delta}_1(\mathbf{T}_1 - \widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_1}) + \widehat{\delta}_2(\mathbf{T}_2 - \widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_2}), \tag{3.10}$$

where $\widehat{\delta}_i$ is the same as the corresponding δ_i in the results of Theorem 3.4, except that $\boldsymbol{\Sigma}_0$ is substituted by $\widetilde{\boldsymbol{\Sigma}}_2$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_i} = \mathbf{P}_1\text{diag}(\mathbf{P}_1^{\mathrm{T}}\mathbf{T}_i\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}}$.

To present further results, define $a_{ij} = \text{tr}[(\mathbf{T}_i - \widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_i})(\mathbf{T}_j - \widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_j})]$ for $i, j = 1, 2$.

Theorem 3.5 *Let Assumptions (A1)-(A4) be satisfied. Assume the split location $m = m(n)$ satisfies the constraint $\sum_{n\geq 1} p(n-m)^{-5} < \infty$ while $p = p(n)$ satisfies $p/n \to c > 0$. If $p^{-1}a_{ii}, p^{-2}(a_{11}a_{22} - a_{12}^2) \nrightarrow 0$ and are finite in probability/almost surely, then we have $\widehat{\delta}_i - \delta_i \to 0$ in probability/almost surely for $i = 1, 2$, where δ_i is as in the result of Theorem 3.4 with $\mathbf{T}_i$ constructed using $\mathbf{Y}_1$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_i} = \mathbf{P}_1\text{diag}(\mathbf{P}_1^{\mathrm{T}}\mathbf{T}_i\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}}$. Furthermore, supposing $\boldsymbol{\Sigma}_0 \neq \sigma^2\mathbf{I}_p$, if we have $\|\mathbf{T}_i - \boldsymbol{\Sigma}_0\| \to 0$ in probability/almost surely while $\|\mathbf{T}_{3-i} - \boldsymbol{\Sigma}_0\| \nrightarrow 0$ but is finite for $i = 1, 2$, then $\widehat{\delta}_i \to 1$, $\widehat{\delta}_{3-i} \to 0$ and $\|\boldsymbol{\Sigma}(\mathbf{P}_1, \mathbf{T}, \widetilde{\boldsymbol{\Sigma}}_2) - \boldsymbol{\Sigma}_0\| \to 0$ in probability/almost surely.*

For data from a factor model, let Assumptions (F1) and (F2) be satisfied. Assume the split location $m = m(n)$ satisfies the constraint $\sum_{n\geq 1} p(n-m)^{-5} < \infty$ while $p = p(n)$ satisfies $p/n \to c > 0$. If $p^{-2}a_{ii}, p^{-4}(a_{11}a_{22} - a_{12}^2) \nrightarrow 0$ and are finite in probability/almost surely, then we have $\widehat{\delta}_i - \delta_i \to 0$ in probability/almost surely for $i = 1, 2$. Furthermore, if $p^{-1}\|\mathbf{T}_i - \boldsymbol{\Sigma}_0\| \to 0$ in probability/almost surely while $p^{-1}\|\mathbf{T}_{3-i} - \boldsymbol{\Sigma}_0\| \nrightarrow 0$ but is finite for $i = 1, 2$, then $\widehat{\delta}_i \to 1$, $\widehat{\delta}_{3-i} \to 0$ and $p^{-1}\|\boldsymbol{\Sigma}(\mathbf{P}_1, \mathbf{T}, \widetilde{\boldsymbol{\Sigma}}_2) - \boldsymbol{\Sigma}_0\| \to 0$ in probability/almost surely.

Theorem 3.5 shows that our estimator is adaptive to the inclusion of two regularized estimators, and favors the correct one asymptotically. This is a very useful property

when we want to take advantages of different regularization methods. See the simulation results in Chapter 3.5 for more details.

Similar to Chapter 3.2.3, we define the ideal estimator as $\mathbf{\Sigma}_{\text{Ideal}} = \mathbf{\Sigma}(\mathbf{P}, \mathbf{T}, \mathbf{\Sigma}_0)$, which is the one in equation (3.9) except that $\mathbf{T}_1$ and $\mathbf{T}_2$ are constructed using $\mathbf{Y}_1$.

Theorem 3.6 *Let Assumptions (A1)-(A4) be satisfied, and* $\mathbf{\Sigma}_0 \neq \sigma^2\mathbf{I}_p$. *Assume the split location* $m = m(n)$ *satisfies the constraints* $m/n \to 1, n - m \to \infty$ *and* $\sum_{n\geq 1} p(n-m)^{-5} < \infty$ *while* $p = p(n)$ *satisfies* $p/n \to c > 0$. *If* $p^{-1}a_{ii}, p^{-2}(a_{11}a_{22} - a_{12}^2) \nrightarrow 0$ *and are finite in probability/almost surely, then* $EL(\mathbf{\Sigma}_0, \mathbf{\Sigma}(\mathbf{P}_1, \mathbf{T}, \widetilde{\mathbf{\Sigma}}_2)) \to 0$ *in probability/almost surely, where* $EL(\mathbf{\Sigma}_0, \widehat{\mathbf{\Sigma}})$ *is as define in equation (3.7) using* $\mathbf{\Sigma}_{\text{Ideal}} = \mathbf{\Sigma}(\mathbf{P}, \mathbf{T}, \mathbf{\Sigma}_0)$.

3.4 Properties of an Averaged Estimator

In this chapter, we illustrate an improved estimator using averaging when two regularized covariance matrices are concerned. The same idea can be applied to the version with one regularized estimator. Since the covariance matrices in equations (3.5) and (3.10) are both constructed from the data $\mathbf{Y}_1$, and the $\mathbf{y}_i$'s are all independent of each other, a new permutation of the data will result in a different $\mathbf{Y}_1$ even if we use the same split location m. (The "permutation" idea is from Lam (2016) with the permutation of $\mathbf{Y}$ and with the same split location m, different subsets $\mathbf{Y}_1$ and $\mathbf{Y}_2$ can be obtained.) Similar to Lam (2016), we permute the data each time, and for the jth permutation where $j = 1, \ldots, M$, we use the resulting data $\mathbf{Y}^{(j)} = (\mathbf{Y}_1^{(j)}, \mathbf{Y}_2^{(j)})$ to construct

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_m^{(j)} &= \mathbf{\Sigma}(\mathbf{P}_{1j}, \mathbf{T}_j, \widetilde{\mathbf{\Sigma}}_2^{(j)}) \\ &= \mathbf{P}_{1j}\text{diag}(\mathbf{P}_{1j}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{P}_{1j})\mathbf{P}_{1j}^{\mathrm{T}} + \widehat{\delta}_{1j}(\mathbf{T}_{1j} - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_{1j}}) + \widehat{\delta}_{2j}(\mathbf{T}_{2j} - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_{2j}}). \qquad (3.11)\end{aligned}$$

In the above, $\widetilde{\mathbf{\Sigma}}_1^{(j)} = m^{-1}\mathbf{Y}_1^{(j)}\mathbf{Y}_1^{(j)\mathrm{T}} = \mathbf{P}_{1j}\mathbf{D}_{1j}\mathbf{P}_{1j}^{\mathrm{T}}$, and $\widetilde{\mathbf{\Sigma}}_2^{(j)} = (n-m)^{-1}\mathbf{Y}_2^{(j)}\mathbf{Y}_2^{(j)\mathrm{T}}$. Each $\mathbf{T}_{ij}$ denotes the ith regularized estimator constructed from $\mathbf{Y}_1^{(j)}$ in the jth permutation, $i = 1, 2,\ j = 1, \ldots, M$. Finally, for $i = 1, 2$ and $j = 1, \ldots, M$, $\widehat{\delta}_{ij}$ is the same as the corresponding δ_i in the result of Theorem 3.4, except that $\mathbf{\Sigma}_0$ is substituted by $\widetilde{\mathbf{\Sigma}}_2^{(j)}$, and $\mathbf{T}_i$ by $\mathbf{T}_{ij}$ with $\widehat{\mathbf{\Sigma}}_{\mathbf{T}_{ij}} = \mathbf{P}_{1j}\text{diag}(\mathbf{P}_{1j}^{\mathrm{T}}\mathbf{T}_{ij}\mathbf{P}_{1j})\mathbf{P}_{1j}^{\mathrm{T}}$. We can then average all these

estimators and arrive at the final estimator

$$\widehat{\mathbf{\Sigma}}_{m,M} = \frac{1}{M}\sum_{j=1}^{M}\widehat{\mathbf{\Sigma}}_m^{(j)}. \tag{3.12}$$

In practice, using $M = 50$ achieves a good trade-off between accuracy and computational efficiency. The resulting estimator is usually much better than using just $M = 1$ alone.

Theorem 3.7 *Let Assumptions (A1)-(A4) be satisfied, and* $\mathbf{\Sigma}_0 \neq \sigma^2\mathbf{I}_p$. *Assume the split location* $m = m(n)$ *satisfies the constraints* $m/n \to 1, n - m \to \infty$ *and* $\sum_{n\geq 1} p(n-m)^{-5} < \infty$ *while* $p = p(n)$ *satisfies* $p/n \to c > 0$. *If the corresponding* $p^{-1}a_{ii}, p^{-2}(a_{11}a_{22} - a_{12}^2) \nrightarrow 0$ *and are finite in probability/almost surely for all* M *permutations, then* $EL(\mathbf{\Sigma}_0, \widehat{\mathbf{\Sigma}}_{m,M}) \leq 0$ *in probability/almost surely.*

Furthermore, if $\sum_{n\geq 1}(n-m)^{-3} < \infty$, *then in probability/almost surely, we have*

$$\frac{1}{p}(\mathrm{tr}(\widehat{\mathbf{\Sigma}}_{m,M}) - \mathrm{tr}(\mathbf{\Sigma}_0)) \to 0, \quad \frac{1}{p}\mathrm{tr}(\widehat{\mathbf{\Sigma}}_{m,M}\mathbf{\Sigma}_0) \geq \lambda_{\min}^2(\mathbf{\Sigma}_0),$$

where $\lambda_{\min}(\mathbf{A})$ *denotes the minimum eigenvalue of a matrix* $\mathbf{A}$. *The first trace property above are true if the data is from a factor model with Assumptions (A1),(A2) replaced by (F1),(F2), and the corresponding* $p^{-2}a_{ii}, p^{-4}(a_{11}a_{22} - a_{12}^2) \nrightarrow 0$ *and are finite in probability/almost surely for all* M *permutations. The second trace property is true if it is modified to*

$$\frac{1}{p^3}\mathrm{tr}(\widehat{\mathbf{\Sigma}}_{m,M}\mathbf{\Sigma}_0) \geq \frac{1}{p^2}\lambda_{\min}^2(\mathbf{\Sigma}_0).$$

The above theorem shows that asymptotic efficiency still holds for the averaged estimator. Moreover, although we cannot prove the asymptotic positive definiteness of the estimator $\widehat{\mathbf{\Sigma}}_{m,M}$ in general, the two trace properties are important characteristics for our estimator. This is because the true covariance matrix $\mathbf{\Sigma}_0$ also satisfies the two trace properties above.

3.4.1 Speed Boosting and Choice of Split Location

The estimator $\widehat{\mathbf{\Sigma}}_{m,M}$ in equation (3.12) involves calculating the regularized covariance matrices $\mathbf{T}_1$ and $\mathbf{T}_2$ M times for a particular split location m. If we search for several split locations (see below), then the computational burden is high even when p is not too large.

To speed up the calculation of the estimator, we concentrate on banding estimator in Bickel and Levina (2008b) and POET in Fan et al. (2013) for $\mathbf{T}_1$ and $\mathbf{T}_2$. POET includes generalized thresholding when set with $K = 0$ factor. To speed up the repeated calculations of the banding estimators, we first find the banding number k chosen by a 5-fold cross-validation (see Bickel and Levina (2008b) for details) when the whole data set $\mathbf{Y}$ is used. Then we use such a k to band each $\mathbf{T}_{1j}$ constructed from the permutated data set $\mathbf{Y}^{(j)}$.

For speeding up the repeated calculations of the POET estimators with K factors, note that an eigen-decomposition of the sample covariance matrix $\widetilde{\mathbf{\Sigma}}_1^{(j)}$ is needed in both the calculation of $\mathbf{P}_{1j}$ and the factor loading matrix in the POET method, and hence one eigen-decomposition is all that is needed in each repeated calculation. It means that we are using the solution $p^{1/2}$ times the matrix of the K column vectors in $\mathbf{P}_{1j}$ corresponding to the K largest eigenvalues of $\widetilde{\mathbf{\Sigma}}_1^{(j)}$ to be the estimated factor loading matrix (see for example Bai and Ng (2002) for more details). We indeed implement this in our codes to minimize the number of eigen-decompositions required.

One major parameter we need to decide is the split location for our estimator. From Theorems 3.3, 3.6 and 3.7, the conditions that have to be satisfied for the split location are that $\sum_{n\geq 1} p(n-m)^{-5} < \infty$, $n - m \to \infty$ and $m/n \to 1$. Coupling with the high-dimensional assumption that $p/n \to c > 0$, one possible solution is to set $m = n - an^{1/2}$. This is indeed a solution adopted by Lam (2016), and we propose a similar criterion for choosing m here:

$$g(m) = \left\| \widehat{\mathbf{\Sigma}}_{m,M} - \frac{1}{M} \sum_{j=1}^{M} \widetilde{\mathbf{\Sigma}}_2^{(j)} \right\|_F^2, \tag{3.13}$$

where $\widetilde{\mathbf{\Sigma}}_2^{(j)}$ is defined in equation (3.11). We choose m that minimizes $g(m)$ above. In practice we search the following 7 split locations for minimizing $g(m)$:

$$m = [2n^{1/2}, 0.2n, 0.4n, 0.6n, 0.8n, n - 2.5n^{1/2}, n - 1.5n^{1/2}].$$

The last two split locations are of the form $n - an^{1/2}$. The four locations $0.2n$ to $0.8n$ are there to accommodate finite sample performance. The split location $2n^{1/2}$ is for the case $\mathbf{\Sigma}_0 = \sigma^2 \mathbf{I}_p$, when all proposed estimators are reduced to NERCOME in Lam (2016) asymptotically, and it is discussed in Lam (2016) explicitly that this case

requires m to be as small as possible while still going to infinity. The order $n^{1/2}$ is a good choice in the end.

3.4.2 Other Practical Concerns

The regularity condition that $\mathbf{T}$ cannot be of the form $\mathbf{PDP}^{\mathrm{T}}$ in Theorem 3.1 does have a bearing in practice. For instance, when $\mathbf{T}$ is a banded or a thresholded estimator, it can be that $\mathbf{T}$ becomes a diagonal matrix and somehow becomes "too close" to the form $\mathbf{PDP}^{\mathrm{T}}$ because the diagonal elements in $\mathbf{T}$ are very similar. This will make $\mathrm{tr}(\mathbf{T} - \widehat{\mathbf{\Sigma}}_{\mathbf{T}})^2$ much smaller than usual, and hence δ will be inflated in magnitude, making the estimator $\mathbf{\Sigma}(\mathbf{P}_1, \mathbf{T}, \widetilde{\mathbf{\Sigma}}_2)$ in equation (3.5) unstable. The same holds true when integrating two regularized estimators $\mathbf{T}_1$ and $\mathbf{T}_2$. For instance if $\mathbf{T}_1$ is banded and $\mathbf{T}_2$ is soft-thresholded, they can both be diagonal with almost identical diagonal elements. This will make the denominator in δ_1 and δ_2 in Theorem 3.4 very small, making the estimator $\mathbf{\Sigma}(\mathbf{P}_1, \mathbf{T}, \widetilde{\mathbf{\Sigma}}_2)$ in equation (3.10) unstable.

In practice, we find that when a weight has a magnitude larger than 3 (can be δ or $1-\delta$, or δ_1,δ_2 or $1-\delta_1-\delta_2$), the corresponding estimator exhibits substantial instability. For a single $\mathbf{T}$, it means that $\mathbf{T}$ is starting to be too close to the form $\mathbf{PDP}^{\mathrm{T}}$, so that the whole problem is approaching the construction of NERCOME in Lam (2016). Hence we impose our estimator to be exactly the NERCOME estimator whenever this happens to protect it from becoming too unstable. For the averaged estimator $\widehat{\mathbf{\Sigma}}_{m,M}$ in equation (3.12), we monitor how many times the estimator is made exactly the NERCOME among the M permutations. Practically, for more accurate estimator, we want at least 30 of the estimators among the M permutations are not made exactly to NERCOME. Suppose there are $M_e \geq 30$ permutations, indexed by the set S so that $M_e = |S|$, where the estimator is not made to NERCOME exactly. Then our average estimator becomes

$$\widehat{\mathbf{\Sigma}}_{m,M} = \frac{1}{M}\sum_{j=1}^{M} \mathbf{P}_{1j}\mathrm{diag}(\mathbf{P}_{1j}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{P}_{1j})\mathbf{P}_{1j}^{\mathrm{T}} + \frac{1}{M_e}\sum_{j\in S}\frac{\mathrm{tr}[(\mathbf{T}_j - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_j})\widetilde{\mathbf{\Sigma}}_2^{(j)}]}{\mathrm{tr}(\mathbf{T}_j - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_j})^2}(\mathbf{T}_j - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_j}). \tag{3.14}$$

If $M_e < 30$, we just take the first term in equation (3.14) above, which is exactly the averaged NERCOME proposed in Lam (2016). This works very well in practice and reduces instability substantially.

We use similar rules for our averaged estimator for integrating $\mathbf{T}_1$ and $\mathbf{T}_2$. For a particular permutation, if any of $\widehat{\delta}_1, \widehat{\delta}_2$ or $1-\widehat{\delta}_1-\widehat{\delta}_2$ are larger than 3 in magnitude, then we discard it. Defining the set S as above, with $M_e = |S|$, if $M_e \geq 30$, the resulting averaged estimator is then

$$\widehat{\mathbf{\Sigma}}_{m,M} = \frac{1}{M}\sum_{j=1}^{M} \mathbf{P}_{1j}\text{diag}(\mathbf{P}_{1j}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{P}_{1j})\mathbf{P}_{1j}^{\mathrm{T}} + \frac{1}{M_e}\sum_{j\in S}\Big(\widehat{\delta}_{1j}(\mathbf{T}_{1j} - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_{1j}}) + \widehat{\delta}_{2j}(\mathbf{T}_{2j} - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_{2j}})\Big). \tag{3.15}$$

If $M_e < 30$, we reduce the above to the one in equation (3.14), getting two averaged estimators, one for $\mathbf{T}_1$ and one for $\mathbf{T}_2$. If say $\mathbf{T}_1$ has $M_e \geq 30$ and $\mathbf{T}_2$ has $M_e < 30$, then we choose the averaged estimator with $\mathbf{T}_1$ and vice versa. If $M_e \geq 30$ for both $\mathbf{T}_1$ and $\mathbf{T}_2$, then it means that originally $\mathbf{T}_1$ and $\mathbf{T}_2$ are perhaps too similar, causing instability in the final estimator. Hence we average the two integrated estimators with $\mathbf{T}_1$ and $\mathbf{T}_2$ in this case. Finally, if both $M_e < 30$, then it is reduced to the averaged NERCOME estimator. We impose all these rules in our simulations and real data analysis.

3.5 Empirical Results

3.5.1 Simulation Experiments

We have three proposed estimators to be compared with other methods in this chapter. They are the estimator integrated with Banding (abbreviated as INT-BAND), the one integrated with POET (abbreviated as INT-POET), and finally the one integrated with both (abbreviated as INT-Double). For POET and related methods, we use $C = 0.5$ throughout. We use the averaged estimator as in equation (3.12), and to speed up the simulations, we use $M = 40$ and search for 4 split locations

$$m = [2n^{1/2}, 0.3n, 0.7n, n - 2n^{1/2}].$$

This way, the time to compute our estimator is cut by half, without losing accuracy and efficiency practically. Other methods to be compared with our estimators include Banding (with banding number chosen by 5-fold cross-validation for each simulated data set) and POET themselves, with POET includes pure adaptive thresholding as a special case. We also compare to three eigenvalues shrinkage methods, namely, the

Nonparametric Eigenvalues-Regularized COvariane Matrix Estimator (NERCOME) from Lam (2016), the nonlinear shrinkage estimator (NONLIN) from Ledoit and Wolf (2012), and the grand average estimator (Grand Avg) from Abadir et al. (2014). Finally, we also compare to the NOVELIST estimator from Huang and Fryzlewicz (2015), which combines the sample covariance matrix with a thresholded estimator. We create 10 different profiles below.

(I) Sparse $\mathbf{\Sigma}_0$ with 20% non-zeros, generated randomly each time, $\mathbf{y}_t \sim N(\mathbf{0}, \mathbf{\Sigma}_0)$.

(II) General non-sparse $\mathbf{\Sigma}_0 = \mathbf{Q}\mathbf{D}\mathbf{Q}^{\mathrm{T}}$, where $\mathbf{Q}$ is an orthogonal matrix generated randomly each time, and $\mathbf{D}$ is diagonal with values 1, 4 , 7 and 10 each appears 25% of times, $\mathbf{y}_t \sim N(\mathbf{0}, \mathbf{\Sigma}_0)$.

(III) (Sparse + Banded) $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2$, where $\mathbf{\Sigma}_1$ is the same as the $\mathbf{\Sigma}_0$ in profile (I), and $\mathbf{\Sigma}_2 = (0.9^{|i-j|})_{1\le i,j\le p}$, $\mathbf{y}_t \sim N(\mathbf{0}, \mathbf{\Sigma}_0)$.

(IV) (Cross) $\mathbf{\Sigma}_0 = \mathbf{Q}\mathrm{diag}(\lambda_1 \mathbf{1}_{\{\lambda_1>0\}}, \ldots, \lambda_p \mathbf{1}_{\{\lambda_p>0\}})\mathbf{Q}^{\mathrm{T}} + \mathbf{I}_p$, where $\mathbf{Q}\mathrm{diag}(\lambda_1, \ldots, \lambda_p)\mathbf{Q}^{\mathrm{T}}$ is the eigen-decomposition of the left-right flipped matrix of $(0.9^{|i-j|})_{1\le i,j\le p}$, $\mathbf{y}_t \sim N(\mathbf{0}, \mathbf{\Sigma}_0)$.

(V) (Factor model) $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{E}$, where $\mathbf{A}$ is $p \times 2$, $\mathbf{X} = (x_1, \ldots, x_n)$ is $2 \times n$, and finally $\mathbf{E} = \mathbf{\Sigma}^{1/2}\mathbf{Z}$, where $\mathbf{\Sigma}$ is the $\mathbf{\Sigma}_0$ in profile (I). The factor loading matrix $\mathbf{A}$, the matrix of factor series $\mathbf{X}$ and $\mathbf{Z}$ are generated each time with independent and identically distributed $N(0, 1)$ elements.

(It-IVt) Same as the corresponding profiles (I) to (IV), except that $\mathbf{Y} = \mathbf{\Sigma}_0^{1/2}\mathbf{Z}$ with $\mathbf{Z}$ having independent and identically distributed t_5 elements.

(Vt) Same as the corresponding profile (V), except that $\mathbf{E} = \mathbf{\Sigma}^{1/2}\mathbf{Z}$ where $\mathbf{Z}$ has independent and identically distributed t_5 elements.

Profile (I) favors POET ($K = 0$), while profile (III) favors banding and profile (V) POET with $K = 2$. Profile (IV) also has sparse $\mathbf{\Sigma}_0$, but the cross-shaped non-zero pattern on $\mathbf{\Sigma}_0$ means it is partly banded as well. Thresholding or banding alone are not able to take full advantage of the structure of $\mathbf{\Sigma}_0$ then. The 5 other profiles are created to test the robustness of our methods to fat-tailed distribution.

We simulate 500 times from the 10 profiles under $n = 200$ and $p = 100, 200, 400$. We use the following 7 loss functions for comparisons:

$$
\begin{aligned}
L_1(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}) &= \|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\|_F, \\
L_2(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}) &= \mathrm{tr}(\mathbf{\Sigma}\widehat{\mathbf{\Sigma}}^{-1}) - \log\det(\mathbf{\Sigma}\widehat{\mathbf{\Sigma}}^{-1}) - p, \\
L_3(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}) &= \|\mathbf{\Sigma}^{-1} - \widehat{\mathbf{\Sigma}}^{-1}\|, \\
L_4(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}) &= \mathrm{tr}(\mathbf{\Sigma}^{-1}\widehat{\mathbf{\Sigma}}) - \log\det(\mathbf{\Sigma}^{-1}\widehat{\mathbf{\Sigma}}) - p, \\
L_5(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}) &= \|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\|, \\
L_6(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}) &= \sum_{i=1}^{p} |\lambda_i(\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}})|, \\
L_7(\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}}) &= \mathrm{tr}(\mathbf{\Sigma} + \widehat{\mathbf{\Sigma}} - 2\mathbf{\Sigma}^{1/2}\widehat{\mathbf{\Sigma}}^{1/2}).
\end{aligned}
$$

The L_1 loss is the Frobenius loss which our estimator is supposed to minimize. The second is the inverse Stein's loss, and the fourth one is the Stein's loss. We also include L_5 the spectral loss and L_6 the nuclear loss, while L_7 is called the Fréchet loss. All of them are non-negative and is 0 when $\widehat{\mathbf{\Sigma}} = \mathbf{\Sigma}$.

From Table 3.1, under profile (I), the INT-BAND and NERCOME have virtually the same performance. This is understandable, since Banding is performing badly, and hence the δ in INT-BAND is small, virtually turning it into NERCOME. POET, being the adaptive soft-thresholding with $K = 0$, is doing well since $\mathbf{\Sigma}_0$ is indeed sparse in profile (I). It is therefore remarkable for INT-POET, which is the combination of NERCOME and POET, to outperform both NERCOME and POET in all 7 losses. Even more remarkable is that INT-Double, being INT-POET combined with Banding, outperforms (albeit only slightly) INT-POET in all 7 losses despite the fact that Banding is not doing good (and not even positive definite in many instances) at all. It would seem that Banding, despite its bad performance, still "contains" some tiny advantages over other estimators, and INT-Double helps extract these tiny advantages out. Incidentally, it outperforms all other methods in all 7 losses, except for NONLIN in L_3.

Within the same profile (I), when $p = n = 200$ (so now $p/n = 1$), it is much more difficult for POET to perform well relative to NERCOME or NONLIN in terms of minimizing the Frobenius loss, hence INT-POET or INT-Double both have weight on POET reduced compared to the case when $p = 100, n = 200$. The results are closer for all the methods as well since the integrated estimators do not have as much advantages

$n = 200$	Profile (I)							
$p = 100$	$\delta(\%)$	L_1	L_2	L_3	L_4	L_5	L_6	L_7
INT-BAND	$-4.7_{(9.6)}$	$37.1_{(1.1)}$	$10.1_{(0.7)}$	$0.8_{(0.1)}$	$15.6_{(2.0)}$	$8.3_{(0.5)}$	$165.4_{(2.5)}$	$39.3_{(1.5)}$
Banding	-	$50.2_{(5.7)}$	$17.2_{(6.8)}$	$1.0_{(0.1)}$	$30.7_{(4.1)}$	$10.7_{(1.9)}$	$190.9_{(9.9)}$	$70.4_{(10.1)}$
INT-POET	$44.7_{(6.9)}$	$35.2_{(1.2)}$	$9.2_{(0.6)}$	$0.8_{(0.1)}$	$14.3_{(1.9)}$	$7.9_{(0.4)}$	$161.0_{(2.8)}$	$35.4_{(1.6)}$
POET	-	$37.6_{(1.5)}$	$10.2_{(0.6)}$	$0.9_{(0.0)}$	$14.2_{(1.3)}$	$8.0_{(0.4)}$	$166.3_{(3.4)}$	$38.9_{(1.6)}$
INT-Double	$\delta_1 : -17.2_{(20.8)}$	$\mathbf{35.0_{(1.3)}}$	$\mathbf{9.1_{(0.6)}}$	$0.8_{(0.1)}$	$\mathbf{14.2_{(1.9)}}$	$\mathbf{7.8_{(0.4)}}$	$\mathbf{160.6_{(2.9)}}$	$\mathbf{35.1_{(1.6)}}$
	$\delta_2 : \quad 47.3_{(7.7)}$							
NERCOME	-	$37.2_{(1.1)}$	$10.1_{(0.7)}$	$0.8_{(0.1)}$	$15.6_{(2.0)}$	$8.3_{(0.5)}$	$165.5_{(2.5)}$	$39.3_{(1.5)}$
NONLIN	-	$36.6_{(1.1)}$	$9.8_{(0.7)}$	$\mathbf{0.7_{(0.1)}}$	$14.2_{(1.2)}$	$8.1_{(0.4)}$	$165.0_{(2.4)}$	$38.1_{(1.2)}$
Grand Avg	-	$37.7_{(1.1)}$	$10.8_{(0.8)}$	$1.0_{(0.0)}$	$20.1_{(2.0)}$	$8.2_{(0.4)}$	$166.6_{(2.4)}$	$42.5_{(1.5)}$
NOVELIST	-	$36.9_{(1.8)}$	$10.3_{(0.8)}$	$0.9_{(0.0)}$	$17.9_{(2.4)}$	$7.9_{(0.5)}$	$165.2_{(4.1)}$	$40.0_{(3.1)}$
$p = 200$	$\delta(\%)$	L_1	L_2	L_3	L_4	L_5	L_6	L_7
INT-BAND	$-4.1_{(6.6)}$	$81.8_{(1.1)}$	$24.3_{(1.3)}$	$0.69_{(0.02)}$	$44.4_{(3.8)}$	$12.5_{(0.4)}$	$413.4_{(2.8)}$	$140.4_{(3.5)}$
Banding	-	$97.9_{(8.6)}$	$36.5_{(47.0)}$	$0.92_{(3.42)}$	$60.1_{(6.3)}$	$14.3_{(1.5)}$	$451.2_{(19.1)}$	$198.8_{(33.9)}$
INT-POET	$22.7_{(4.6)}$	$80.2_{(1.2)}$	$23.5_{(1.2)}$	$0.69_{(0.02)}$	$43.0_{(3.7)}$	$12.2_{(0.4)}$	$409.5_{(3.0)}$	$135.4_{(3.5)}$
POET	-	$92.0_{(1.9)}$	$30.5_{(1.1)}$	$\mathbf{0.64_{(0.02)}}$	$\mathbf{39.5_{(2.5)}}$	$13.6_{(0.5)}$	$438.6_{(4.5)}$	$163.9_{(3.5)}$
INT-Double	$\delta_1 : -12.9_{(12.4)}$	$\mathbf{80.0_{(1.2)}}$	$\mathbf{23.5_{(1.2)}}$	$0.69_{(0.02)}$	$42.8_{(3.7)}$	$\mathbf{12.2_{(0.4)}}$	$\mathbf{409.1_{(3.1)}}$	$\mathbf{134.9_{(3.5)}}$
	$\delta_2 : \quad 24.2_{(4.8)}$							
NERCOME	-	$81.8_{(1.1)}$	$24.3_{(1.3)}$	$0.69_{(0.02)}$	$44.4_{(3.8)}$	$12.5_{(0.4)}$	$413.4_{(2.8)}$	$140.4_{(3.5)}$
NONLIN	-	$81.0_{(1.1)}$	$24.0_{(1.5)}$	$0.69_{(0.02)}$	$43.4_{(3.4)}$	$12.2_{(0.4)}$	$413.0_{(2.8)}$	$138.3_{(3.5)}$
Grand Avg	-	$82.5_{(1.1)}$	$24.7_{(1.3)}$	$0.70_{(0.02)}$	$48.1_{(4.2)}$	$12.3_{(0.3)}$	$414.5_{(2.8)}$	$145.0_{(3.9)}$
NOVELIST	-	$83.3_{(1.6)}$	$38.6_{(2.4)}$	$0.72_{(0.02)}$	$45.9_{(4.0)}$	$12.4_{(0.4)}$	$418.6_{(3.9)}$	$145.8_{(4.7)}$

Table 3.1 Mean and standard deviation (in bracket) of different losses for different methods: Profile (I)

* For POET and related methods, $K = 0$ is used. For banding, we omit the instances where the estimator is not positive definite when $p = 100$ in profile (I) and report the 20% trimmed mean together with its interquartile range/1.349 for L_2, L_3 and L_7. Gray cells indicate the minimum among all methods.

$n = 200$	Profile (It)							
$p = 200$	$\delta(\%)$	L_1	L_2	L_3	L_4	L_5	L_6	L_7
INT-BAND	$-4.6_{(6.2)}$	$81.8_{(1.1)}$	$24.4_{(1.2)}$	$0.69_{(0.02)}$	$44.6_{(4.0)}$	$12.5_{(0.4)}$	$413.5_{(2.9)}$	$140.6_{(3.5)}$
Banding	-	$113.9_{(24.1)}$	$38.1_{(10.8)}$	$1.03_{(0.02)}$	$70.0_{(17.7)}$	$21.1_{(9.1)}$	$481.1_{(43.2)}$	$239.7_{(64.3)}$
INT-POET	$20.0_{(4.7)}$	$80.6_{(1.1)}$	$23.8_{(1.1)}$	$0.69_{(0.02)}$	$43.4_{(3.9)}$	$12.3_{(0.4)}$	$410.5_{(2.9)}$	$136.7_{(3.6)}$
POET	-	$95.2_{(3.3)}$	$31.7_{(1.0)}$	$\mathbf{0.65_{(0.02)}}$	$\mathbf{41.2_{(2.5)}}$	$17.6_{(6.5)}$	$443.3_{(5.0)}$	$172.3_{(5.7)}$
INT-Double	$\delta_1 : -13.9_{(13.1)}$	$\mathbf{80.1_{(1.2)}}$	$\mathbf{23.6_{(1.1)}}$	$0.69_{(0.02)}$	$42.9_{(4.0)}$	$\mathbf{12.2_{(0.4)}}$	$\mathbf{409.3_{(3.3)}}$	$\mathbf{135.1_{(3.9)}}$
	$\delta_2 : 24.3_{(6.4)}$							
NERCOME	-	$81.8_{(1.1)}$	$24.4_{(1.2)}$	$0.69_{(0.02)}$	$44.6_{(4.0)}$	$12.5_{(0.4)}$	$413.6_{(2.8)}$	$140.8_{(3.6)}$
NONLIN	-	$81.5_{(1.8)}$	$24.4_{(1.8)}$	$0.69_{(0.06)}$	$43.2_{(3.2)}$	$13.1_{(4.3)}$	$414.1_{(3.1)}$	$139.3_{(3.7)}$
Grand Avg	-	$82.5_{(1.1)}$	$24.8_{(1.2)}$	$0.70_{(0.02)}$	$48.6_{(4.0)}$	$12.3_{(0.4)}$	$414.7_{(2.8)}$	$145.6_{(3.7)}$
NOVELIST	-	$88.8_{(3.3)}$	$40.5_{(2.2)}$	$0.72_{(0.02)}$	$48.1_{(4.1)}$	$16.7_{(6.9)}$	$429.0_{(4.6)}$	$159.9_{(6.4)}$
$n = 200$	Profile (I)							
$p = 400$	$\delta(\%)$	L_1	L_2	L_3	L_4	L_5	L_6	L_7
INT-BAND	$-2.5_{(4.4)}$	$171.2_{(1.1)}$	$52.1_{(2.0)}$	$0.50_{(0.01)}$	$99.0_{(6.7)}$	$17.9_{(0.3)}$	$1007.6_{(3.5)}$	$438.9_{(8.5)}$
Banding	-	$192.2_{(17.3)}$	$65.0_{(19.8)}$	$0.50_{(0.01)}$	$117.4_{(12.0)}$	$19.7_{(2.2)}$	$1065.7_{(45.1)}$	$546.5_{(97.7)}$
INT-POET	$8.5_{(1.5)}$	$170.5_{(1.1)}$	$51.7_{(2.0)}$	$0.50_{(0.01)}$	$98.2_{(6.6)}$	$17.8_{(0.3)}$	$1005.4_{(3.5)}$	$435.3_{(8.3)}$
POET	-	$221.5_{(2.8)}$	$96.2_{(2.3)}$	$\mathbf{0.47_{(0.01)}}$	$107.7_{(4.6)}$	$24.9_{(0.8)}$	$1143.3_{(6.9)}$	$670.6_{(7.4)}$
INT-Double	$\delta_1 : -7.3_{(6.2)}$	$170.4_{(1.1)}$	$51.7_{(2.0)}$	$0.50_{(0.01)}$	$\mathbf{98.1_{(6.6)}}$	$17.8_{(0.3)}$	$\mathbf{1005.2_{(3.5)}}$	$\mathbf{434.9_{(8.3)}}$
	$\delta_2 : 9.1_{(1.6)}$							
NERCOME	-	$171.2_{(1.1)}$	$52.1_{(2.0)}$	$0.50_{(0.01)}$	$99.0_{(6.7)}$	$17.9_{(0.3)}$	$1007.6_{(3.5)}$	$438.9_{(8.5)}$
NONLIN	-	$\mathbf{170.2_{(1.1)}}$	$\mathbf{51.7_{(2.0)}}$	$0.50_{(0.01)}$	$99.2_{(6.5)}$	$\mathbf{17.6_{(0.3)}}$	$1006.2_{(3.5)}$	$436.3_{(8.3)}$
Grand Avg	-	$172.1_{(1.1)}$	$52.5_{(2.1)}$	$0.50_{(0.01)}$	$102.1_{(6.8)}$	$17.8_{(0.3)}$	$1009.2_{(3.5)}$	$445.6_{(8.8)}$
NOVELIST	-	$174.9_{(1.2)}$	$68.6_{(2.9)}$	$0.51_{(0.01)}$	$101.6_{(6.5)}$	$18.0_{(0.3)}$	$1020.6_{(3.9)}$	$456.7_{(8.5)}$

Table 3.2 Mean and standard deviation (in bracket) of different losses for different methods: Profile (I) and (It)

* For POET and related methods, $K = 0$ is used. For banding, we omit the instances where the estimator is not positive definite when $p = 200$ in profile (It) and report the 20% trimmed mean together with its interquartile range/1.349 for L_2, L_3 and L_7. Gray cells indicate the minimum among all methods.

over other estimators as before as a result of POET doing worse than before relative to other estimators (except for L_3). This trend continues as we increase p from 200 to 400 in Table 3.2, when the ratio p/n becomes 2, which is difficult for regularized estimators to perform well with this relatively small n. The weight on POET for INT-POET and INT-Double decrease further, as POET itself is performing even worse than Banding on average. Hence both integrated estimators, albeit still better than NERCOME itself, is much closer to NERCOME in performance over all losses. The results for $p = 200$ under profile (It) is also included in Table 3.2, and the pattern is similar to that for profile (I) when $p = 200$. Profile (It) for $p = 100$ and $p = 400$ are similar to the respective profile (I) for $p = 100$ and $p = 400$ too and we omit the results to save space.

From Table 3.3, the performance of the integrated estimators under profile (II) or (IIt) are virtually the same as NERCOME, which is to be expected since $\mathbf{\Sigma}_0$ has no sparsity or banded structure to be exploit, so that the corresponding weights for Banding and POET are small around 0. As such, NONLIN shows slightly better performance on average compared to NERCOME under profile (II) or (IIt), which is also documented in Lam (2016). From Table 3.4, profile (III) and (IIIt) feature a $\mathbf{\Sigma}_0$ which is constructed by adding a sparse and a banded covariance matrix together, so that $\mathbf{\Sigma}_0$ itself is partly banded at best, and further from the diagonal it is roughly approximately sparse. Hence it is expected for the integrated estimators to put more emphasis on POET with $K = 0$, and less on Banding. This is exactly the case from Table 3.2, where the weight for banding is on average positive but in fact fluctuates around 0 for all integrated estimators. The emphasis on POET is much larger, although similar to profile (I) or (It) in Table 3.1, the weight decreases in general as p increases from 200 to 400. Under profile (III) or (IIIt), the integrated estimators, especially INT-Double, has an edge over all other estimators on average.

Table 3.5 shows the results under profile (IV) and (IVt). Because of the cross non-zero pattern on $\mathbf{\Sigma}_0$, banding can result in gross deviation from the true values in the off-diagonal entries as long as the banding number is not very large. It is no surprise that Banding is performing very badly. INT-BAND is performing worse than NERCOME on L_3, but is still close to NERCOME on other losses. Looking at the corresponding weight, Banding still exploits and takes advantages of the special structure of $\mathbf{\Sigma}_0$, but definitely not much. POET itself is also not performing particularly well, but INT-POET, having taken advantages of the sparsity of $\mathbf{\Sigma}_0$ from POET itself, has performed well in all the losses. INT-Double suffers from the bad performance

$n = 200$	Profile (II)							
$p = 100$	$\delta(\%)$	L_1	L_2	L_3	L_4	L_5	L_6	L_7
INT-BAND	$0.4_{(3.7)}$	$23.1_{(0.2)}$	$15.5_{(0.3)}$	$0.72_{(0.02)}$	$23.5_{(1.1)}$	$5.0_{(0.2)}$	$132.0_{(0.8)}$	$28.2_{(0.6)}$
Banding	-	$36.5_{(2.6)}$	$42.9_{(25.7)}$	$18.53_{(22.8)}$	$54.2_{(14.9)}$	$8.2_{(2.6)}$	$162.0_{(2.2)}$	$69.6_{(4.4)}$
INT-POET	$-0.1_{(5.2)}$	$23.1_{(0.2)}$	$15.5_{(0.3)}$	$0.72_{(0.02)}$	$23.5_{(1.1)}$	$5.0_{(0.2)}$	$132.0_{(0.8)}$	$28.2_{(0.6)}$
POET	-	$28.3_{(0.2)}$	$23.5_{(0.4)}$	$0.78_{(0.01)}$	$41.0_{(0.8)}$	$5.7_{(0.2)}$	$146.3_{(0.8)}$	$44.7_{(0.6)}$
INT-Double	$\delta_1: 0.4_{(3.7)}$	$23.1_{(0.3)}$	$15.5_{(0.3)}$	$0.72_{(0.02)}$	$23.5_{(1.1)}$	$5.0_{(0.2)}$	$132.0_{(0.8)}$	$28.2_{(0.6)}$
	$\delta_2: -0.2_{(5.2)}$							
NERCOME	-	$23.1_{(0.2)}$	$15.5_{(0.3)}$	$0.72_{(0.02)}$	$23.5_{(1.1)}$	$5.0_{(0.2)}$	$132.0_{(0.8)}$	$28.2_{(0.6)}$
NONLIN	-	$\mathbf{22.7_{(0.2)}}$	$\mathbf{14.9_{(0.4)}}$	$\mathbf{0.71_{(0.03)}}$	$\mathbf{21.3_{(0.7)}}$	$\mathbf{4.9_{(0.2)}}$	$\mathbf{130.4_{(0.8)}}$	$\mathbf{27.0_{(0.5)}}$
Grand Avg	-	$23.9_{(0.2)}$	$18.3_{(0.2)}$	$0.74_{(0.01)}$	$34.0_{(0.8)}$	$4.9_{(0.1)}$	$136.7_{(0.8)}$	$33.6_{(0.5)}$
NOVELIST	-	$25.9_{(0.2)}$	$52.4_{(1.0)}$	$0.89_{(0.00)}$	$37.7_{(1.0)}$	$5.3_{(0.1)}$	$141.0_{(0.8)}$	$38.4_{(0.5)}$
$n = 200$	Profile (IIt)							
$p = 200$	$\delta(\%)$	L_1	L_2	L_3	L_4	L_5	L_6	L_7
INT-BAND	$-1.4_{(4.2)}$	$39.1_{(0.3)}$	$45.9_{(0.6)}$	$0.78_{(0.01)}$	$89.0_{(6.5)}$	$5.5_{(0.2)}$	$294.5_{(2.3)}$	$89.1_{(2.6)}$
Banding	-	$63.6_{(16.9)}$	$107.3_{(94.5)}$	$58.08_{(76.45)}$	$140.7_{(16.4)}$	$12.1_{(6.6)}$	$355.4_{(25.6)}$	$211.5_{(87.0)}$
INT-POET	$-1.3_{(3.5)}$	$39.2_{(0.3)}$	$45.9_{(0.6)}$	$0.78_{(0.01)}$	$89.1_{(6.5)}$	$5.5_{(0.2)}$	$294.5_{(2.3)}$	$89.1_{(2.5)}$
POET	-	$47.0_{(1.8)}$	$61.3_{(0.7)}$	$0.81_{(0.01)}$	$105.3_{(2.5)}$	$7.6_{(3.6)}$	$316.5_{(1.5)}$	$119.6_{(2.7)}$
INT-Double	$\delta_1: -1.5_{(4.5)}$	$39.1_{(0.3)}$	$45.9_{(0.6)}$	$\mathbf{0.78_{(0.01)}}$	$89.1_{(6.5)}$	$5.5_{(0.2)}$	$294.5_{(2.3)}$	$89.1_{(2.6)}$
	$\delta_2: -0.8_{(3.7)}$							
NERCOME	-	$39.1_{(0.3)}$	$\mathbf{45.9_{(0.6)}}$	$0.78_{(0.01)}$	$89.0_{(6.5)}$	$5.5_{(0.2)}$	$294.5_{(2.3)}$	$89.1_{(2.6)}$
NONLIN	-	$\mathbf{38.9_{(2.0)}}$	$46.0_{(1.9)}$	$0.88_{(0.12)}$	$\mathbf{82.1_{(3.5)}}$	$6.2_{(3.4)}$	$\mathbf{290.2_{(1.1)}}$	$\mathbf{86.4_{(2.4)}}$
Grand Avg	-	$40.3_{(0.2)}$	$48.8_{(0.2)}$	$0.78_{(0.01)}$	$106.8_{(2.3)}$	$\mathbf{5.3_{(0.1)}}$	$300.8_{(1.1)}$	$98.5_{(0.9)}$
NOVELIST	-	$42.8_{(1.9)}$	$90.0_{(1.2)}$	$0.88_{(0.00)}$	$104.8_{(2.7)}$	$7.0_{(3.8)}$	$305.0_{(1.4)}$	$104.2_{(2.7)}$

Table 3.3 Mean and standard deviation (in bracket) of different losses for different methods: Profile (II) and (IIt)

* For POET and related methods, $K = 0$ is used. For banding, we omit each instance where the estimator is not positive definite in all profiles and report the 20% trimmed mean together with its interquartile range/1.349 for all losses. For NONLIN under profile (IIt) with $p = 200$, we do the same for L_2 and L_3, when some instances of NONLIN are near singular. Gray cells indicate the minimum among all methods.

$n = 200$	Profile (IIIt)						
$p = 200$	$\delta(\%)$	L_1	L_2	L_3	L_4	L_5	L_6	L_7
INT-BAND	$6.9_{(6.8)}$	$89.6_{(1.2)}$	$23.1_{(1.1)}$	$0.47_{(0.03)}$	$37.9_{(3.4)}$	**$18.2_{(0.7)}$**	$426.4_{(2.8)}$	$148.2_{(3.9)}$
Banding	-	$119.2_{(16.8)}$	$40.8_{(14.4)}$	$0.84_{(0.04)}$	$57.4_{(8.5)}$	$21.9_{(5.6)}$	$492.5_{(34.3)}$	$259.9_{(66.5)}$
INT-POET	$15.4_{(4.2)}$	$89.2_{(1.2)}$	$22.9_{(1.1)}$	$0.47_{(0.03)}$	$37.3_{(3.4)}$	$18.6_{(0.6)}$	**$424.5_{(3.0)}$**	$146.3_{(4.0)}$
POET	-	$106.7_{(18.5)}$	$31.7_{(1.1)}$	**$0.43_{(0.03)}$**	$38.1_{(3.6)}$	$22.5_{(24.6)}$	$462.3_{(4.9)}$	$191.0_{(18.3)}$
INT-Double	$\delta_1 : 3.1_{(10.4)}$	**$89.0_{(1.2)}$**	**$22.8_{(1.1)}$**	$0.47_{(0.03)}$	$37.2_{(3.4)}$	$18.3_{(0.7)}$	$424.5_{(2.9)}$	**$145.8_{(4.0)}$**
	$\delta_2 : 14.4_{(5.3)}$							
NERCOME	-	$90.0_{(1.1)}$	$23.2_{(1.1)}$	$0.47_{(0.03)}$	$38.0_{(3.4)}$	$18.7_{(0.6)}$	$426.4_{(2.8)}$	$148.9_{(3.9)}$
NONLIN	-	$91.3_{(20.8)}$	$23.0_{(1.6)}$	$0.47_{(0.05)}$	**$37.1_{(3.3)}$**	$20.5_{(25.9)}$	$428.2_{(3.2)}$	$148.8_{(17.6)}$
Grand Avg	-	$90.5_{(1.1)}$	$23.5_{(1.2)}$	$0.48_{(0.03)}$	$40.8_{(3.5)}$	$18.6_{(0.6)}$	$427.4_{(2.9)}$	$153.1_{(4.2)}$
NOVELIST	-	$98.8_{(19.2)}$	$40.3_{(2.4)}$	$0.49_{(0.03)}$	$41.2_{(4.2)}$	$22.4_{(24.8)}$	$443.6_{(4.9)}$	$170.6_{(18.5)}$
$p = 400$	$\delta(\%)$	L_1	L_2	L_3	L_4	L_5	L_6	L_7
INT-BAND	$3.8_{(4.6)}$	$180.1_{(1.2)}$	$49.9_{(1.2)}$	$0.37_{(0.02)}$	$87.0_{(4.0)}$	**$22.5_{(0.5)}$**	$1026.5_{(3.2)}$	$448.1_{(6.1)}$
Banding	-	$231.5_{(42.5)}$	$74.6_{(18.5)}$	$0.37_{(0.02)}$	$113.6_{(13.1)}$	$30.5_{(9.2)}$	$1143.0_{(66.1)}$	$667.7_{(133.9)}$
INT-POET	$7.1_{(2.1)}$	$179.8_{(1.1)}$	$49.7_{(1.2)}$	$0.37_{(0.02)}$	$86.5_{(4.0)}$	$22.7_{(0.4)}$	$1025.0_{(3.2)}$	$446.3_{(6.1)}$
POET	-	$237.5_{(3.3)}$	$94.8_{(2.1)}$	**$0.34_{(0.02)}$**	$102.2_{(3.4)}$	$32.6_{(9.7)}$	$1176.1_{(5.4)}$	$715.0_{(11.4)}$
INT-Double	$\delta_1 : 1.2_{(6.3)}$	$179.7_{(1.2)}$	$49.6_{(1.2)}$	$0.37_{(0.02)}$	**$86.5_{(4.0)}$**	$22.5_{(0.5)}$	**$1025.0_{(3.3)}$**	**$445.7_{(6.1)}$**
	$\delta_2 : 6.7_{(2.5)}$							
NERCOME	-	$180.4_{(1.1)}$	$49.9_{(1.2)}$	$0.37_{(0.02)}$	$87.0_{(4.0)}$	$22.7_{(0.4)}$	$1026.6_{(3.2)}$	$448.8_{(6.1)}$
NONLIN	-	**$179.7_{(1.3)}$**	**$49.6_{(1.2)}$**	$0.37_{(0.02)}$	$87.0_{(3.7)}$	$22.7_{(2.7)}$	$1027.9_{(3.5)}$	$447.0_{(6.7)}$
Grand Avg	-	$181.1_{(1.1)}$	$50.2_{(1.2)}$	$0.37_{(0.02)}$	$90.2_{(4.1)}$	$22.6_{(0.4)}$	$1028.0_{(3.4)}$	$455.5_{(6.8)}$
NOVELIST	-	$194.7_{(4.5)}$	$70.3_{(2.1)}$	$0.37_{(0.02)}$	$93.8_{(4.4)}$	$29.1_{(10.5)}$	$1064.5_{(9.6)}$	$508.1_{(19.2)}$

Table 3.4 Mean and standard deviation (in bracket) of different losses for different methods: Profile (IIIt)

* For POET and related methods, $K = 0$ is used. For banding, we omit each instance where the estimator is not positive definite in all profiles and report the 20% trimmed mean together with its interquartile range/1.349 for all losses. Gray cells indicate the minimum among all methods.

$n = 200$	Profile (IV)							
$p = 100$	$\delta(\%)$	L_1	L_2	L_3	L_4	L_5	L_6	L_7
INT-BAND	$-25.4_{(14.1)}$	$8.5_{(0.7)}$	$3.9_{(0.2)}$	$0.56_{(0.23)}$	$3.4_{(0.2)}$	$5.2_{(1.0)}$	$46.0_{(1.3)}$	$4.0_{(0.3)}$
Banding	-	$11.7_{(1.0)}$	$71.0_{(59.2)}$	$48.23_{(45.37)}$	$35.2_{(3.2)}$	$5.0_{(0.9)}$	$82.3_{(0.8)}$	$18.7_{(1.7)}$
INT-POET	$38.4_{(7.4)}$	$8.3_{(0.7)}$	$\mathbf{3.4_{(0.2)}}$	$\mathbf{0.35_{(0.03)}}$	$\mathbf{3.1_{(0.1)}}$	$5.2_{(1.0)}$	$45.5_{(0.9)}$	$\mathbf{3.6_{(0.3)}}$
POET	-	$11.2_{(0.8)}$	$15.1_{(1.3)}$	$2.12_{(0.24)}$	$12.6_{(0.8)}$	$8.0_{(1.1)}$	$69.2_{(0.9)}$	$9.0_{(0.4)}$
INT-Double	$\delta_1 : -24.8_{(13.9)}$	$\mathbf{8.3_{(0.7)}}$	$3.5_{(0.3)}$	$0.57_{(0.29)}$	$3.2_{(0.2)}$	$5.2_{(1.1)}$	$45.9_{(0.9)}$	$3.6_{(0.3)}$
	$\delta_2 : \quad 38.3_{(7.2)}$							
NERCOME	-	$8.5_{(0.7)}$	$3.9_{(0.2)}$	$0.42_{(0.03)}$	$3.3_{(0.1)}$	$5.2_{(1.0)}$	$45.6_{(1.3)}$	$4.0_{(0.3)}$
NONLIN	-	$8.3_{(0.6)}$	$3.8_{(0.2)}$	$0.42_{(0.02)}$	$3.4_{(0.2)}$	$\mathbf{4.9_{(0.9)}}$	$\mathbf{45.3_{(1.1)}}$	$3.9_{(0.3)}$
Grand Avg	-	$8.6_{(0.7)}$	$3.8_{(0.2)}$	$0.41_{(0.02)}$	$3.3_{(0.1)}$	$5.4_{(1.1)}$	$47.7_{(1.1)}$	$4.0_{(0.3)}$
NOVELIST	-	$10.5_{(0.7)}$	$225.1_{(14.4)}$	$6.56_{(0.62)}$	$23.4_{(7.2)}$	$4.9_{(1.1)}$	$77.3_{(4.0)}$	$13.5_{(3.1)}$
$n = 200$	Profile (IVt)							
$p = 200$	$\delta(\%)$	L_1	L_2	L_3	L_4	L_5	L_6	L_7
INT-BAND	$-17.5_{(30.2)}$	$15.3_{(0.6)}$	$12.5_{(0.8)}$	$1.65_{(1.40)}$	$10.5_{(1.1)}$	$7.7_{(0.8)}$	$100.9_{(4.3)}$	$12.6_{(0.7)}$
Banding	-	$22.8_{(1.0)}$	$> 10^3$	$> 10^3$	$151.4_{(42.0)}$	$8.3_{(1.6)}$	$187.4_{(12.9)}$	$71.8_{(11.7)}$
INT-POET	$62.4_{(5.7)}$	$14.1_{(0.7)}$	$\mathbf{10.3_{(0.6)}}$	$0.48_{(0.08)}$	$9.9_{(0.7)}$	$7.4_{(0.9)}$	$106.3_{(1.8)}$	$\mathbf{10.5_{(0.7)}}$
POET	-	$17.3_{(0.8)}$	$45.9_{(3.4)}$	$3.91_{(0.55)}$	$33.4_{(1.9)}$	$9.5_{(0.8)}$	$146.1_{(1.5)}$	$23.0_{(1.0)}$
INT-Double	$\delta_1 : -16.6_{(25.6)}$	$\mathbf{14.0_{(0.7)}}$	$10.8_{(0.8)}$	$1.20_{(0.95)}$	$10.3_{(0.9)}$	$7.4_{(0.9)}$	$107.0_{(2.0)}$	$10.6_{(0.7)}$
	$\delta_2 : \quad 61.0_{(6.1)}$							
NERCOME	-	$15.3_{(0.6)}$	$11.9_{(0.4)}$	$0.50_{(0.04)}$	$9.8_{(0.6)}$	$7.8_{(0.8)}$	$100.0_{(4.5)}$	$12.4_{(0.6)}$
NONLIN	-	$15.1_{(0.7)}$	$11.6_{(0.4)}$	$0.50_{(0.04)}$	$10.4_{(0.9)}$	$\mathbf{7.1_{(1.0)}}$	$\mathbf{99.1_{(3.0)}}$	$12.4_{(0.8)}$
Grand Avg	-	$15.5_{(0.7)}$	$11.8_{(0.4)}$	$\mathbf{0.48_{(0.04)}}$	$\mathbf{9.8_{(0.7)}}$	$8.1_{(0.8)}$	$106.9_{(3.2)}$	$12.5_{(0.7)}$
NOVELIST	-	$21.9_{(0.7)}$	$389.6_{(20.5)}$	$6.59_{(0.58)}$	$229.0_{(28.9)}$	$7.5_{(1.2)}$	$191.7_{(5.2)}$	$74.4_{(6.7)}$

Table 3.5 Mean and standard deviation (in bracket) of different losses for different methods.

* For POET and related methods, $K = 0$ is used for profile (IV) and (IVt). For banding, we omit each instance where the estimator is not positive definite in both profiles and report the 20% trimmed mean together with its interquartile range/1.349 for all losses. Same for INT-Band and INT-Double under profile (IVt) with L_2 and L_3 losses. Gray cells indicate the minimum among all methods.

$n = 200$	Profile (V)							
$p = 200$	$\delta(\%)$	L_1	L_2	L_3	L_4	L_5	L_6	L_7
INT-BAND	$0.1_{(5.9)}$	$138.2_{(8.0)}$	$25.6_{(1.3)}$	$0.66_{(0.04)}$	$45.6_{(3.9)}$	$79.2_{(12.0)}$	$432.0_{(3.0)}$	$172.2_{(5.1)}$
Banding	-	$229.4_{(7.3)}$	$> 10^3$	$> 10^2$	$153.1_{(24.0)}$	$76.8_{(10.5)}$	$631.0_{(9.9)}$	$787.8_{(21.7)}$
INT-POET	$26.4_{(5.3)}$	$137.3_{(8.1)}$	$\mathbf{24.9_{(1.2)}}$	$0.66_{(0.04)}$	$44.2_{(3.7)}$	$79.2_{(12.0)}$	$\mathbf{428.5_{(3.2)}}$	$\mathbf{167.6_{(5.1)}}$
POET	-	$139.2_{(5.5)}$	$26.8_{(1.1)}$	$\mathbf{0.64_{(0.04)}}$	$\mathbf{42.2_{(3.3)}}$	$\mathbf{70.5_{(8.2)}}$	$437.0_{(3.7)}$	$175.6_{(4.6)}$
INT-Double	$\delta_1: 0.0_{(5.7)}$ $\delta_2: 26.4_{(5.3)}$	$137.3_{(8.1)}$	$24.9_{(1.2)}$	$0.66_{(0.04)}$	$44.2_{(3.7)}$	$79.2_{(12.0)}$	$428.5_{(3.2)}$	$167.6_{(5.1)}$
NERCOME	-	$138.2_{(8.0)}$	$25.6_{(1.3)}$	$0.66_{(0.04)}$	$45.6_{(3.9)}$	$79.3_{(12.0)}$	$432.0_{(3.0)}$	$172.2_{(5.1)}$
NONLIN	-	$\mathbf{136.0_{(6.7)}}$	$25.3_{(1.3)}$	$0.66_{(0.04)}$	$45.0_{(3.7)}$	$73.0_{(10.4)}$	$432.2_{(3.0)}$	$170.3_{(4.7)}$
Grand Avg	-	$139.5_{(8.4)}$	$26.0_{(1.3)}$	$0.67_{(0.04)}$	$49.6_{(4.3)}$	$81.8_{(12.3)}$	$433.1_{(3.0)}$	$177.5_{(5.3)}$
NOVELIST	-	$229.3_{(7.2)}$	$255.8_{(37.7)}$	$0.71_{(0.04)}$	$231.0_{(3.1)}$	$78.9_{(12.1)}$	$629.4_{(9.0)}$	$801.6_{(24.4)}$
$n = 200$	Profile (Vt)							
$p = 400$	$\delta(\%)$	L_1	L_2	L_3	L_4	L_5	L_6	L_7
INT-BAND	$0.0_{(1.5)}$	$305.1_{(13.0)}$	$55.3_{(1.9)}$	$0.49_{(0.01)}$	$102.5_{(6.5)}$	$175.7_{(20.9)}$	$1038.9_{(3.5)}$	$533.6_{(11.3)}$
Banding	-	$616.6_{(15.5)}$	-	-	-	$187.6_{(44.8)}$	$1706.8_{(80.7)}$	-
INT-POET	$7.6_{(3.0)}$	$304.8_{(13.0)}$	$55.1_{(1.9)}$	$0.49_{(0.01)}$	$\mathbf{102.0_{(6.5)}}$	$175.7_{(20.9)}$	$\mathbf{1037.5_{(3.5)}}$	$\mathbf{531.2_{(11.4)}}$
POET	-	$319.7_{(9.0)}$	$71.7_{(1.8)}$	$\mathbf{0.47_{(0.01)}}$	$103.7_{(5.6)}$	$\mathbf{148.2_{(12.4)}}$	$1110.2_{(6.0)}$	$642.9_{(14.0)}$
INT-Double	$\delta_1: -0.1_{(1.5)}$ $\delta_2: 7.5_{(3.0)}$	$304.8_{(13.0)}$	$55.1_{(1.9)}$	$0.49_{(0.01)}$	$102.0_{(6.5)}$	$175.7_{(20.9)}$	$1037.5_{(3.5)}$	$531.2_{(11.4)}$
NERCOME	-	$305.0_{(13.0)}$	$55.3_{(1.9)}$	$0.49_{(0.01)}$	$102.5_{(6.5)}$	$175.7_{(20.9)}$	$1038.8_{(3.5)}$	$533.5_{(11.3)}$
NONLIN	-	$\mathbf{300.1_{(10.1)}}$	$\mathbf{55.0_{(1.9)}}$	$0.49_{(0.01)}$	$102.3_{(6.3)}$	$159.3_{(17.8)}$	$1040.4_{(4.1)}$	$531.2_{(11.3)}$
Grand Avg	-	$308.2_{(14.6)}$	$55.7_{(2.0)}$	$0.49_{(0.01)}$	$106.9_{(6.8)}$	$182.1_{(21.5)}$	$1040.3_{(3.6)}$	$542.3_{(12.7)}$
NOVELIST	-	$623.5_{(13.3)}$	-	$0.52_{(0.02)}$	-	$179.9_{(20.5)}$	$1738.9_{(15.5)}$	-

Table 3.6 Mean and standard deviation (in bracket) of different losses for different methods: Profile (V) and (Vt)

* For POET and related methods, $K = 2$ is used for profile (V) and (Vt). For banding, we omit each instance where the estimator is not positive definite in all profiles and report the 20% trimmed mean together with its interquartile range/1.349 for all losses. Under profile (Vt) with $p = 400$, almost all instances of Banding and NOVELIST are non-positive definite, and the corresponding L_2, L_3, L_4 and L_7 are not reported. Gray cells indicate the minimum among all methods.

of Banding itself, and is not performing as good as INT-POET in L_3, although it still helps INT-Double to achieve the smallest Frobenius loss L_1 overall. NONLIN and Grand Avg also have good performance overall, especially when $p = 200$. Under profile (V) in Table 3.6, with the correct number of factors specified, POET performs well overall, but INT-POET still outperforms POET in L_2, L_6 and L_7. The weight on Banding is virtually 0, which should be the case ideally, so that INT-Double has virtually the same performance as INT-POET. The same goes for profile (Vt) with $p = 400$, where the weight on Banding is virtually 0 for INT-BAND and INT-Double. The weight on POET has decreased comparing to profile (V) at $p = 200$, so that NONLIN and Grand Avg have closer performance to the integrated estimators. Again, this reflects the increased difficulties for POET to achieve a good performance as $p/n = 2$ with n relatively small at 200.

3.5.2 Forecasting the Number of Phone Calls

We are interested in forecasting the number of phone calls for a call center. The data is used in Huang et al. (2006) and Bickel and Levina (2008b), and is re-analyzed in Lam (2016). Phone calls to a call center are recorded from 7am to midnight everyday in 2002, except for weekends, holiday and when equipments are malfunctioning, leaving $n = 239$ days of calls in total. A 17-hour recording period is divided into 10-minute intervals each day, resulting in 102 intervals. Let N_{ij} be the number of calls at the jth interval on the ith day, $i = 1, \ldots, 239$, $j = 1, \ldots, 102$. The transformation $y_{ij} = (N_{ij} + 1/4)^{1/2}$ is applied to bring the data closer to normal (see Huang et al. (2006) for more details).

We consider predicting the number of phone calls in the latter half of the day for the last 29-day period of the data using the data from the first half of the day. Lam (2016) has illustrated empirically that the last 29-day period of the data is particularly difficult to forecast, and we bring in our proposed estimators to compare to other methods used in the simulations to see if our proposed estimators can improve forecasting. In details, let $\mathbf{y}_i = ((\mathbf{y}_i^{(1)})^{\mathrm{T}}, (\mathbf{y}_i^{(2)})^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{y}_i^{(1)} = (y_{i,1}, \ldots, y_{i,51})^{\mathrm{T}}$ and $\mathbf{y}_i^{(2)} = (y_{i,52}, \ldots, y_{i,102})^{\mathrm{T}}$. Let $\boldsymbol{\mu}_j = E(\mathbf{y}_i^{(j)})$ for $j = 1, 2$, and $\boldsymbol{\Sigma}_{jk} = \mathrm{cov}(\mathbf{y}_i^{(j)}\mathbf{y}_i^{(k)})$, $1 \le j, k \le 2$. We use the best linear predictor

$$\hat{\mathbf{y}}_i^{(2)} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_i^{(1)} - \boldsymbol{\mu}_1)$$

for predicting the number of calls in the second half of the day. We need to estimate $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{y}_i)$ using past data. We consider 30,60,90,120,150,180 and 210 days of training

data prior to the 29-day period we want to forecast, and for each training data set, we compare all the methods used in the simulations in Chapter 3.5.1, plus sample covariance, pure adaptive soft thresholding (SOFT, i.e., POET with $K = 0$), integrated estimator with soft thresholding (INT-SOFT), and integrated estimator with banding and soft-thresholding (INT-bandsoft). For POET, we consider $K = 1$ to explore if some parts of the data exhibits factor model structure. For banding, we consider the banding of the modified Cholesky factor instead of direct banding of the covariance matrix (see Bickel and Levina (2008b) for more details), since the components in $\mathbf{y}_i$ are time-ordered, so the dependence on distant past can be weak, meaning the modified Cholesky factor can be naturally banded.

For $k = 30, 60, \ldots, 210$, we compare for different methods the absolute forecast errors

$$E_k = \frac{1}{51}\sum_{j=52}^{102} E_{k,j}, \quad \text{where } E_{k,j} = \frac{1}{29}\sum_{r=211}^{239} |\hat{y}_{rj,k}^{(2)} - y_{rj}^{(2)}|, \tag{3.16}$$

where $\hat{y}_{rj,k}^{(2)}$ is the forecast on day r and time interval j, using k prior days for constructing an estimator for $\mathbf{\Sigma}$. Except for SOFT, POET and NONLIN, we repeatedly estimate $\mathbf{\Sigma}$ for 80 times and average the forecast results to arrive at $\hat{y}_{rj,k}^{(2)}$ to reduce the variability from data splitting, and the choice of banding number for banding or the choice of thresholding parameter for NOVELIST by cross-validation.

From Table 3.7, the best forecasts come from soft-thresholding when we use more data ($k \geq 150$) in estimating $\mathbf{\Sigma}$, followed closely by INT-Double and INT-bandsoft when using $k \leq 90$. Banding is good in general for larger k, which makes sense since a larger k means some components in $\mathbf{y}_i$ are further apart in time, so that the modified Cholesky factor is more banded. INT-SOFT and INT-bandsoft perform well in general for each particular k, and are in general better than the nonlinear shrinkage methods like NERCOME, NONLIN and Grand Avg. It indicates that, despite the bad performance for banding and soft-thresholding when $k \leq 60$, there are still advantages from the banding and sparse assumptions, and the corresponding integrated estimators help squeeze them out. Certainly, when k gets larger, it becomes apparent that $\mathbf{\Sigma}$ becomes sparser and with more banded structure, and banding and soft-thresholding themselves fare better.

The integrated estimators INT-SOFT and INT-bandsoft are also better than the NOVELIST in general, even when NERCOME and sample covariance are performing

	Training data size (days prior to the forecast period)						
	30	60	90	120	150	180	210
Sample	-	$2.94_{(0.83)}$	$1.71_{(0.50)}$	$1.60_{(0.47)}$	$1.59_{(0.52)}$	$1.60_{(0.51)}$	$1.53_{(0.49)}$
NERCOME	$1.45_{(0.50)}$	$1.51_{(0.47)}$	$1.59_{(0.51)}$	$1.58_{(0.52)}$	$1.59_{(0.53)}$	$1.60_{(0.55)}$	$1.53_{(0.51)}$
NONLIN	$1.48_{(0.52)}$	$1.50_{(0.48)}$	$1.60_{(0.51)}$	$1.60_{(0.52)}$	$1.61_{(0.54)}$	$1.60_{(0.55)}$	$1.53_{(0.51)}$
Grand Avg	$1.46_{(0.50)}$	$1.50_{(0.47)}$	$1.59_{(0.51)}$	$1.59_{(0.52)}$	$1.60_{(0.53)}$	$1.60_{(0.55)}$	$1.54_{(0.51)}$
NOVELIST	$1.61_{(0.50)}$	$1.72_{(0.51)}$	$1.64_{(0.50)}$	$1.57_{(0.48)}$	$1.57_{(0.52)}$	$1.59_{(0.51)}$	$1.52_{(0.48)}$
Banding	$1.88_{(0.54)}$	$1.74_{(0.51)}$	$1.50_{(0.47)}$	$\mathbf{1.47_{(0.47)}}$	$1.51_{(0.51)}$	$1.54_{(0.52)}$	$1.44_{(0.46)}$
INT-BAND	$1.43_{(0.49)}$	$1.50_{(0.47)}$	$1.58_{(0.50)}$	$1.58_{(0.52)}$	$1.59_{(0.53)}$	$1.60_{(0.55)}$	$1.53_{(0.51)}$
POET	$1.75_{(0.55)}$	$1.61_{(0.49)}$	$1.71_{(0.54)}$	$1.68_{(0.54)}$	$1.69_{(0.56)}$	$1.68_{(0.56)}$	$1.54_{(0.51)}$
INT-POET	$1.45_{(0.50)}$	$1.50_{(0.48)}$	$1.55_{(0.49)}$	$1.55_{(0.51)}$	$1.58_{(0.53)}$	$1.59_{(0.55)}$	$1.51_{(0.50)}$
INT-Double	$\mathbf{1.42_{(0.49)}}$	$1.51_{(0.48)}$	$1.55_{(0.50)}$	$1.56_{(0.51)}$	$1.58_{(0.53)}$	$1.59_{(0.55)}$	$1.51_{(0.50)}$
SOFT	$8.88_{(4.30)}$	$4.81_{(1.99)}$	$2.08_{(0.80)}$	$1.57_{(0.46)}$	$\mathbf{1.39_{(0.43)}}$	$\mathbf{1.42_{(0.45)}}$	$\mathbf{1.37_{(0.42)}}$
INT-SOFT	$1.47_{(0.50)}$	$1.47_{(0.38)}$	$1.44_{(0.42)}$	$1.48_{(0.46)}$	$1.52_{(0.50)}$	$1.55_{(0.53)}$	$1.50_{(0.49)}$
INT-bandsoft	$1.46_{(0.49)}$	$\mathbf{1.45_{(0.39)}}$	$\mathbf{1.44_{(0.42)}}$	$1.48_{(0.46)}$	$1.52_{(0.50)}$	$1.55_{(0.53)}$	$1.50_{(0.49)}$

Table 3.7 Mean and standard deviation (in bracket) of the $E_{k,j}$'s defined in (3.16) for different methods.

* Sample covariance does not have a stable $\mathbf{\Sigma}_{11}^{-1}$ for $k = 30$, and the corresponding results are omitted. Gray cells indicate the minimum for a particular training data size k.

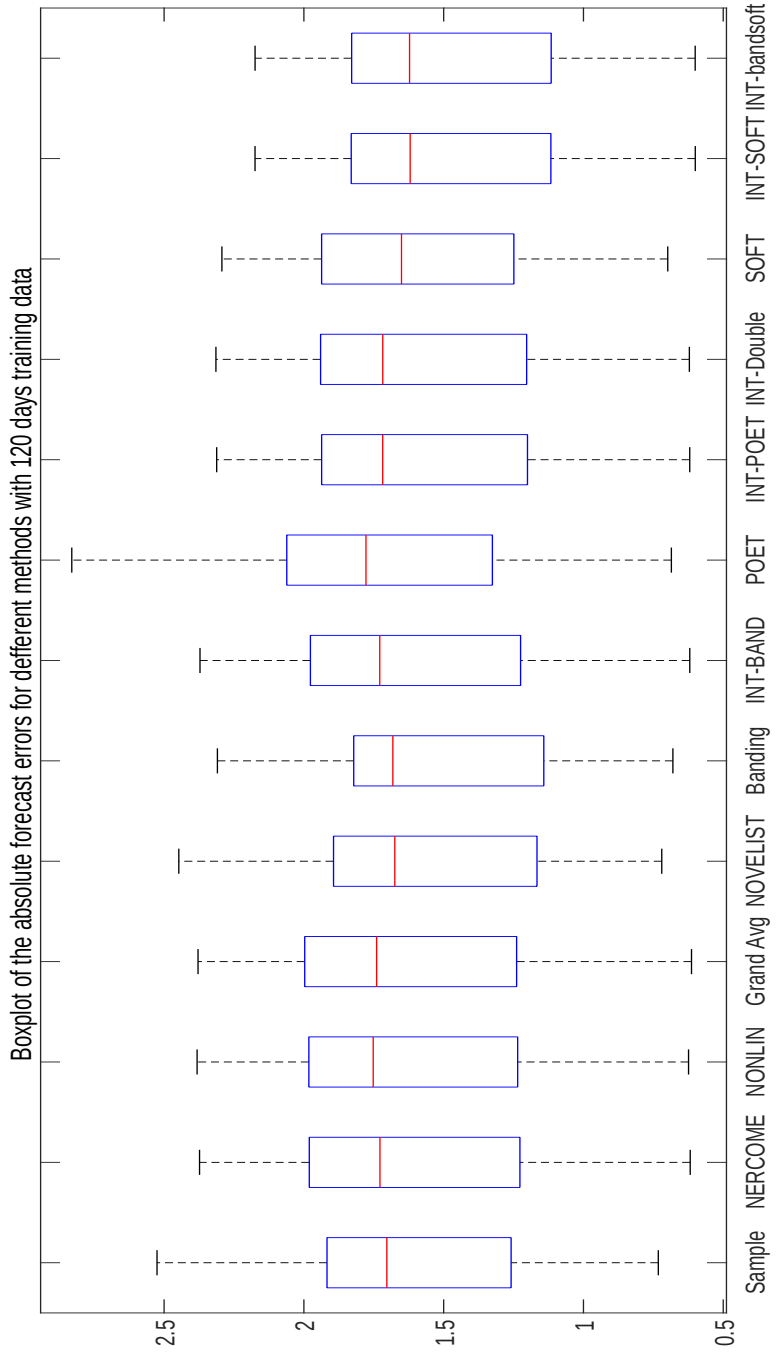


Fig. 3.1 Boxplot of the absolute forecast errors $E_{k,j}$ in (3.16) for different methods with 120 days training data.

at the same level when $k \geq 150$. It means that the integrated estimators are better in extracting the advantages from the sparsity of $\mathbf{\Sigma}$ on average in this study.

To better take a colse look at the comparisions of different methods, we present an example boxplot of all estimators' absolute errors with $k = 120$ in Figure 3.1, where the proposed estimators INT-SOFT and INT-bandsoft outperform almost all other methods with a lower mean as well as standard deviation.

Remark 3.2 *Due to the uncerntainty of the positive definiteness of proposed estimators, same as most of the covariance matrix estimators, we do not provide the stock market experiments in the following empirical study. The percentage of having positive definite estimator varies with different dataset and different choices of the regularized estimators. For example, if we estimate based on a 4-week training window with 26 NYSE stocks' daily closing log-price (the detailed background is similar to Chapter 4.5.4), 100% positive definiteness can be obtained by INT-BAND, INT-Double, INT-SOFT, INT-bandsoft, INT-bandcrc, INT-softcrc and INT-bandsoftcrc (where "crc" represents "Grand Avg"). However, INT-POET achieves 97.96% positive definiteness rate while INT-poetcrc and INT-doublecrc only have the rates as 26.53% and 73.47%. As a result, the choice of regularized would be important. We would suggest for a careful choices of regularized estimators. For non positive definite estimator, we propose to diagonalise the proposed estimator and replace any eigenvalues that fall under a certain small positive threshold by the value of that threshold.*

3.6 Proof of Theorems

Proof of Theorem 3.1. Define $R(\delta, \mathbf{D})$ as the squared Frobenius norm of $\mathbf{\Sigma}(\delta, \mathbf{D}) - \mathbf{\Sigma}_0$ in equation (3.2) in the chapter. Then

$$\begin{aligned} R(\delta, \mathbf{D}) &= \left\| (1-\delta)\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}} + \delta\mathbf{T} - \mathbf{\Sigma}_0 \right\|_F^2 \\ &= \left\| (1-\delta)(\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}} - \mathbf{\Sigma}_0) + \delta(\mathbf{T} - \mathbf{\Sigma}_0) \right\|_F^2 \\ &= (1-\delta)^2 \left\| \mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}} - \mathbf{\Sigma}_0 \right\|_F^2 + 2\delta(1-\delta)\mathrm{tr}[(\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}} - \mathbf{\Sigma}_0)(\mathbf{T} - \mathbf{\Sigma}_0)] \\ &\quad + \delta^2 \left\| \mathbf{T} - \mathbf{\Sigma}_0 \right\|_F^2. \end{aligned} \tag{3.17}$$

First, we calculate the solution for $\mathbf{D} = \text{diag}(d_1, \ldots, d_p)$ to the minimization problem in terms of δ. For each $j = 1, \ldots, p$, we have for $\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_p)$,

$$\begin{aligned}
\frac{\partial R(\delta, \mathbf{D})}{\partial d_j} &= (1-\delta)^2 \frac{\partial}{\partial d_j}(\text{tr}(\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}}\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}}) - 2\text{tr}(\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0)) \\
&\quad + 2\delta(1-\delta)\frac{\partial}{\partial d_j}(\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}}(\mathbf{T} - \boldsymbol{\Sigma}_0)) \\
&= (1-\delta)^2 \frac{\partial}{\partial d_j}(\sum_{i=1}^{p} d_i^2 - 2\sum_{i=1}^{p} d_i \mathbf{p}_i^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_i) + 2\delta(1-\delta)\frac{\partial}{\partial d_j}(\sum_{i=1}^{p} d_i \mathbf{p}_i^{\mathrm{T}}(\mathbf{T} - \boldsymbol{\Sigma}_0)\mathbf{p}_i) \\
&= (1-\delta)^2(2d_j - 2\mathbf{p}_j^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_j) + 2\delta(1-\delta)\mathbf{p}_j^{\mathrm{T}}(\mathbf{T} - \boldsymbol{\Sigma}_0)\mathbf{p}_j \\
&= 2[(1-\delta)^2 d_j - (1-\delta)^2\mathbf{p}_j^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_j + \delta(1-\delta)\mathbf{p}_j^{\mathrm{T}}(\mathbf{T} - \boldsymbol{\Sigma}_0)\mathbf{p}_j].
\end{aligned}$$

Set $\frac{\partial R(\delta, \mathbf{D})}{\partial d_j} = 0$, then since we assumed $\delta \neq 1$,

$$\begin{aligned}
d_j &= \frac{(1-\delta)^2\mathbf{p}_j^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_j - \delta(1-\delta)\mathbf{p}_j^{\mathrm{T}}(\mathbf{T} - \boldsymbol{\Sigma}_0)\mathbf{p}_j}{(1-\delta)^2} \\
&= \mathbf{p}_j^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_j - \frac{\delta}{1-\delta}(\mathbf{p}_j^{\mathrm{T}}\mathbf{T}\mathbf{p}_j - \mathbf{p}_j^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_j) \\
&= \frac{1}{1-\delta}\mathbf{p}_j^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_j - \frac{\delta}{1-\delta}\mathbf{p}_j^{\mathrm{T}}\mathbf{T}\mathbf{p}_j, \quad \text{so that} \\
\mathbf{D} &= \frac{1}{1-\delta}\text{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P}) - \frac{\delta}{1-\delta}\text{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{T}\mathbf{P}). \qquad (3.18)
\end{aligned}$$

This proves the first part of Theorem 1. For the second part of the proof, we substitute $\mathbf{D}$ from (3.18) into the original function $R(\delta, \mathbf{D})$ in (3.17). Hence

$$\begin{aligned}
R(\delta, \mathbf{D}) = R(\delta) &= \left\|\mathbf{P}\text{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}} - \delta\mathbf{P}\text{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{T}\mathbf{P})\mathbf{P}^{\mathrm{T}} + \delta\mathbf{T} - \boldsymbol{\Sigma}_0\right\|_F^2 \\
&= \left\|\mathbf{P}\text{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}} - \boldsymbol{\Sigma}_0\right\|_F^2 + \delta^2\left\|\hat{\boldsymbol{\Sigma}}_{\mathbf{T}} - \mathbf{T}\right\|_F^2 \\
&\quad - 2\delta\text{tr}[(\mathbf{P}\text{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}} - \boldsymbol{\Sigma}_0)(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}} - \mathbf{T})]. \\
\frac{\partial R(\delta)}{\partial \delta} &= 2\delta\left\|\hat{\boldsymbol{\Sigma}}_{\mathbf{T}} - \mathbf{T}\right\|_F^2 - 2\text{tr}[(\mathbf{P}\text{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}} - \boldsymbol{\Sigma}_0)(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}} - \mathbf{T})] \\
&= 2[\delta\text{tr}(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}} - \mathbf{T})^2 - \text{tr}(\boldsymbol{\Sigma}_0(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}} - \mathbf{T}))], \qquad (3.19)
\end{aligned}$$

where $\hat{\boldsymbol{\Sigma}}_{\mathbf{T}} = \mathbf{P}\text{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{T}\mathbf{P})\mathbf{P}$, and we arrive at the equality in (3.19) since

$$\begin{aligned}
&\mathrm{tr}(\mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}}(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}}-\mathbf{T})) \\
&= \mathrm{tr}(\mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}}\mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{T}\mathbf{P})\mathbf{P}^{\mathrm{T}}) - \mathrm{tr}(\mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}}\mathbf{T}) \\
&= \mathrm{tr}(\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{T}\mathbf{P})) - \mathrm{tr}(\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}}\mathbf{T}\mathbf{P}) = 0. \qquad (3.20)
\end{aligned}$$

Setting $\frac{\partial R(\delta)}{\partial \delta} = 0$, we have the result as stated in Theorem 3.1, since $\mathrm{tr}(\mathbf{T}-\hat{\boldsymbol{\Sigma}}_{\mathbf{T}})^2 \neq 0$ as $\mathbf{T}$ is not of the form $\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}}$ for some diagonal matrix $\mathbf{D}$. □

To prove Theorem 3.2, we need to state a lemma first, which closely resembles Lemma 1 of Lam (2016).

Lemma 3.1 *Let Assumptions (A1) be satisfied. If the split location m is such that $\sum_{n\geq 1} p(n-m)^{-5} < \infty$, then we have*

$$\max_{1\leq i\leq p}\left|\frac{\mathbf{q}_i^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{q}_i - \mathbf{q}_i^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{q}_i}{\mathbf{q}_i^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{q}_i}\right| \to 0$$

almost surely, where $\mathbf{q}_1, \ldots, \mathbf{q}_p$ are unit vectors independent of the data $\mathbf{Y}_2$. The same holds true of the data is from a factor model, with Assumption (F1) satisfied together with $\sum_{n\geq 1} p(n-m)^{-5} < \infty$.

The proof of this lemma is exactly the same as that for Lemma 1 of Lam (2016), the only difference is the substitution of $\mathbf{p}_{1i}$ there by $\mathbf{q}_i$ here.

Proof of Theorem 3.2. With $\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}}$ being real symmetric, assume that $\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}} = \mathbf{Q}\mathbf{D}_{\mathbf{T}}\mathbf{Q}^{\mathrm{T}}$, where $\mathbf{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_p)$ is orthogonal and $\mathbf{D}_{\mathbf{T}} = \mathrm{diag}(d_1, \ldots, d_p)$ is diagonal. The $\mathbf{q}_i$'s are independent of $\mathbf{Y}_2$ since $\mathbf{T}$ is constructed from the data $\mathbf{Y}_1$. Then with Assumptions (A1) to (A2) assuming the data is not from a factor model,

$$\begin{aligned}
|\hat{\delta} - \delta| &= \frac{|\mathrm{tr}[(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})(\widetilde{\boldsymbol{\Sigma}}_2 - \boldsymbol{\Sigma}_0)]|}{\mathrm{tr}(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2} = \frac{\left|\sum_{i=1}^p d_i(\mathbf{q}_i^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{q}_i - \mathbf{q}_i\boldsymbol{\Sigma}_0\mathbf{q}_i)\right|}{\mathrm{tr}(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2} \\
&\leq \frac{\left(p^{-1}\sum_{i=1}^p d_i^2\right)^{1/2}\left(p^{-1}\sum_{i=1}^p(\mathbf{q}_i^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{q}_i - \mathbf{q}_i\boldsymbol{\Sigma}_0\mathbf{q}_i)^2\right)^{1/2}}{p^{-1}\mathrm{tr}(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2} \\
&\leq \frac{\max_{1\leq i\leq p}\left|\frac{\mathbf{q}_i^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{q}_i - \mathbf{q}_i^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{q}_i}{\mathbf{q}_i^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{q}_i}\right|\left(p^{-1}\sum_{i=1}^p(\mathbf{q}_i^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{q}_i)^2\right)^{1/2}}{\left(p^{-1}\mathrm{tr}(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2\right)^{1/2}} \to 0 \qquad (3.21)
\end{aligned}$$

in probability/almost surely by Lemma 3.1, the fact that $(p^{-1}\sum_{i=1}^{p}(\mathbf{q}_i^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{q}_i)^2)^{1/2} \leq \lambda_{\max}(\boldsymbol{\Sigma}_0) < \infty$ by Assumption (A2), and finally by the assumption that $p^{-1}\mathrm{tr}(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2 \nrightarrow 0$ in probability/almost surely. This proves $\widehat{\delta} - \delta \to 0$ in probability/almost surely.

Now consider

$$\begin{aligned}\widehat{\delta} &= \frac{\mathrm{tr}[(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})\boldsymbol{\Sigma}_0]}{\mathrm{tr}(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2} = \frac{\mathrm{tr}[(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})(\boldsymbol{\Sigma}_0 - \mathbf{T})] + \mathrm{tr}[(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})\mathbf{T}]}{\mathrm{tr}(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2} \\ &= 1 + \frac{\mathrm{tr}[(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})(\boldsymbol{\Sigma}_0 - \mathbf{T})]}{\mathrm{tr}(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2},\end{aligned}$$

where the last line follows since $\mathrm{tr}[(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}}] = 0$. But with the assumption $\|\mathbf{T} - \boldsymbol{\Sigma}_0\| \to 0$ in probability/almost surely, then

$$\begin{aligned}\left|\frac{\mathrm{tr}[(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})(\boldsymbol{\Sigma}_0 - \mathbf{T})]}{\mathrm{tr}(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2}\right| &\leq \frac{\left(p^{-1}\mathrm{tr}(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2\right)^{1/2}\left(p^{-1}\mathrm{tr}(\mathbf{T} - \boldsymbol{\Sigma}_0)^2\right)^{1/2}}{p^{-1}\mathrm{tr}(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2} \\ &\leq \frac{\|\mathbf{T} - \boldsymbol{\Sigma}_0\|}{\left(p^{-1}\mathrm{tr}(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2\right)^{1/2}} \to 0\end{aligned}$$

in probability/almost surely by the assumption that $p^{-1}\mathrm{tr}(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2 \nrightarrow 0$ in probability/almost surely. This proves that $\widehat{\delta} \to 1$ in probability/almost surely. With this, then

$$\begin{aligned}&\|\boldsymbol{\Sigma}(\mathbf{P}_1, \mathbf{T}, \widetilde{\boldsymbol{\Sigma}}_2) - \boldsymbol{\Sigma}_0\| \\ &= \|\mathbf{P}_1\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}} + \widehat{\delta}(\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}}) - \boldsymbol{\Sigma}_0\| \\ &\leq \|\mathbf{P}_1\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}}\| + \|\mathbf{T} - \boldsymbol{\Sigma}_0\| + |\widehat{\delta} - 1| \cdot \|\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}}\| \\ &= \max_{1\leq i\leq p} |\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\mathbf{T}\mathbf{p}_{1i}| + \|\mathbf{T} - \boldsymbol{\Sigma}_0\| + |\widehat{\delta} - 1| \cdot \|\mathbf{T} - \widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}}\| \\ &\leq \max_{1\leq i\leq p}\left|\frac{\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_{1i}}\right| \cdot \|\boldsymbol{\Sigma}_0\| + \max_{1\leq i\leq p} |\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{T}\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_{1i}| \\ &\quad + \|\mathbf{T} - \boldsymbol{\Sigma}_0\| + 2|\widehat{\delta} - 1| \cdot \|\mathbf{T}\| \\ &\leq \max_{1\leq i\leq p}\left|\frac{\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_{1i}}\right| \cdot \|\boldsymbol{\Sigma}_0\| + 2\|\mathbf{T} - \boldsymbol{\Sigma}_0\| + 2|\widehat{\delta} - 1| \cdot \|\mathbf{T}\| \to 0\end{aligned}$$

in probability/almost surely, where the last line follows from Lemma 1 of Lam (2016), Assumption (A2) that $\|\mathbf{\Sigma}_0\|$ is finite, and the assumption that $\|\mathbf{T}\|$ is finite in probability/almost surely.

For data from a factor model, almost every details of the proofs are exactly the same as before, except that we now have $p^{-2}\text{tr}(\mathbf{T} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}})^2 \nrightarrow 0$ while $p^{-1}\|\mathbf{\Sigma}_0\| = O(1)$ and $p^{-1}\|\mathbf{T}\|$ are finite in probability/almost surely. This completes the proof of the theorem. □

To prove Theorem 3.3, we need to state and prove another lemma first.

Lemma 3.2 *Let Assumptions (A1) to (A4) be satisfied, and* $\mathbf{\Sigma}_0 \neq \sigma^2\mathbf{I}_p$. *Assume the split location* $m = m(n)$ *satisfies the constraints* $m/n \to 1$, $n - m \to \infty$ *and* $\sum_{n\geq 1} p(n-m)^{-5} < \infty$ *while* $p = p(n)$ *satisfies* $p/n \to c > 0$. *Then we have* $\|\mathbf{P}_1 - \mathbf{P}\| \to 0$ *almost surely. In particular, assuming* $\|\mathbf{T}\|$ *finite in probability/almost surely, we have in probability/almost surely,*

$$\frac{1}{p}\sum_{i=1}^{p} |\mathbf{p}_{1i}^{\text{T}}\mathbf{T}\mathbf{p}_{1i} - \mathbf{p}_i^{\text{T}}\mathbf{T}\mathbf{p}_i| \to 0, \quad \frac{1}{p}\sum_{i=1}^{p} |\mathbf{p}_{1i}^{\text{T}}\mathbf{\Sigma}_0\mathbf{p}_{1i} - \mathbf{p}_i^{\text{T}}\mathbf{\Sigma}_0\mathbf{p}_i| \to 0.$$

Proof of Lemma 3.2. The setting is the same as that in Lemma S.4 of Lam (2016). As such, for λ_i and λ_{1i} being the eigenvalues of $\widetilde{\mathbf{\Sigma}} = n^{-1}\mathbf{Y}\mathbf{Y}^{\text{T}}$ and $\widetilde{\mathbf{\Sigma}}_1 = m^{-1}\mathbf{Y}_1\mathbf{Y}_1^{\text{T}}$ respectively (with corresponding eigenvectors $\mathbf{p}_i$ and $\mathbf{p}_{1i}$), we have for any continuous function $g(\cdot)$ over the positive real line,

$$p^{-1}\sum_{i=1}^{p} g(\mathbf{p}_i^{\text{T}}\mathbf{\Sigma}_0\mathbf{p}_i)\mathbf{1}_{\{\lambda_i \leq x\}} \overset{a.s.}{\to} \int_{\infty}^{x} g(\delta(\lambda))dF(\lambda), \tag{3.22}$$

where $F(\cdot)$ is such that

$$F_p(\lambda) = p^{-1}\sum_{i=1}^{p} \mathbf{1}_{\{\lambda_i \leq \lambda\}} \overset{a.s.}{\to} F(\lambda).$$

equation (3.22) is the same as equation (S.3) in Lam (2016). Please refer to equation (2.7) of Lam (2016) for the definition of $\delta(\cdot)$. Similarly, we have

$$p^{-1}\sum_{i=1}^{p} g(\mathbf{p}_{1i}^{\text{T}}\mathbf{\Sigma}_0\mathbf{p}_{1i})\mathbf{1}_{\{\lambda_{1i} \leq x\}} \overset{a.s.}{\to} \int_{\infty}^{x} g(\delta_1(\lambda))dF_1(\lambda), \tag{3.23}$$

where $\delta_1(\cdot)$ is as in (2.9) of Lam (2016), and $F_1(\cdot)$ is such that

$$F_{1p}(\lambda) = p^{-1} \sum_{i=1}^{p} \mathbf{1}_{\{\lambda_{1i} \leq \lambda\}} \xrightarrow{a.s.} F_1(\lambda).$$

But since $m/n \to 1$, we have $p/m, p/n$ both go to the same limit $c > 0$. Theorem 4.1 of Bai and Silverstein (2010) tells us then both F_p and F_{1p} converge to the same limit almost surely under Assumptions (A1) to (A4). Hence $F = F_1$ almost surely, implying $\delta_1(\cdot) = \delta(\cdot)$ almost surely. Setting $g \equiv 1$, we see that λ_{1i} and λ_i are almost surely the same.

Then using Theorem 3 of Ledoit and Péché (2011), we can arrive at that both $\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{v}_j$ and $\mathbf{p}_i^{\mathrm{T}}\mathbf{v}_j$ for $i, j = 1, \ldots, p$ are almost surely the same as a function $\varphi(\lambda_i, \tau_j)$ uniformly for each i and j (which depends on the same constant c where $p/m, p/n \to c > 0$ since $m/n \to 1$), where $\mathbf{v}_j$ is the eigenvector of $\mathbf{\Sigma}_0$ corresponding to the j-th largest eigenvalue τ_j of $\mathbf{\Sigma}_0$. Hence

$$\max_{1 \leq i,j \leq p} |(\mathbf{p}_{1i} - \mathbf{p}_i)^{\mathrm{T}}\mathbf{v}_j| \xrightarrow{a.s.} 0.$$

It means that we must have $\max_{1 \leq i \leq p} \|\mathbf{p}_{1i} - \mathbf{p}_i\| \xrightarrow{a.s.} 0$. Hence $\|\mathbf{P}_1 - \mathbf{P}\| \xrightarrow{a.s.} 0$ follows. With this, then

$$\frac{1}{p}\sum_{i=1}^{p} |\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{T}\mathbf{p}_{1i} - \mathbf{p}_i^{\mathrm{T}}\mathbf{T}\mathbf{p}_i| = \frac{1}{p}\sum_{i=1}^{p} |(\mathbf{p}_{1i} - \mathbf{p}_i)^{\mathrm{T}}\mathbf{T}(\mathbf{p}_{1i} + \mathbf{p}_i)| \leq 2\|\mathbf{T}\| \max_{1 \leq i \leq p} \|\mathbf{p}_{1i} - \mathbf{p}_i\| \to 0$$

in probability/almost surely, since $\|\mathbf{T}\|$ is finite in probability/almost surely. Replacing $\mathbf{T}$ by $\mathbf{\Sigma}_0$ proves the almost sure convergence of the other by Assumption (A2) that $\|\mathbf{\Sigma}_0\| = O(1)$. □

Proof of Theorem 3.3. Recall the notations $\widehat{\mathbf{\Sigma}}_{\mathbf{T}} = \mathbf{P}\text{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{T}\mathbf{P})\mathbf{P}^{\mathrm{T}}$ and $\widetilde{\mathbf{\Sigma}}_{\mathbf{T}} = \mathbf{P}_1\text{diag}(\mathbf{P}_1^{\mathrm{T}}\mathbf{T}\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}}$. Consider

$$\begin{aligned}
\frac{1}{p}\left\|\mathbf{\Sigma}_{\text{Ideal}} - \mathbf{\Sigma}_0\right\|_F^2 &= \frac{1}{p}\left\|\mathbf{P}\text{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}} - \mathbf{\Sigma}_0 + \frac{\text{tr}[(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})\mathbf{\Sigma}_0]}{\text{tr}(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})^2}(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})\right\|_F^2 \\
&= \frac{1}{p}\left\|\mathbf{P}\text{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}} - \mathbf{\Sigma}_0\right\|_F^2 + \frac{1}{p}\frac{\text{tr}^2[(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})\mathbf{\Sigma}_0]}{\text{tr}^2(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})^2}\text{tr}(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})^2 \\
&\quad + \frac{2}{p}\text{tr}[(\mathbf{P}\text{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}} - \mathbf{\Sigma}_0)(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})]\frac{\text{tr}[(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})\mathbf{\Sigma}_0]}{\text{tr}(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})^2} \\
&= \frac{1}{p}\left\|\text{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{P}) - \mathbf{P}^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{P}\right\|_F^2 - \frac{1}{p}\frac{\text{tr}^2[(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})\mathbf{\Sigma}_0]}{\text{tr}(\mathbf{T}-\widehat{\mathbf{\Sigma}}_{\mathbf{T}})^2}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\frac{1}{p}\left\|\mathbf{\Sigma}(\mathbf{P}_1,\mathbf{T},\widetilde{\mathbf{\Sigma}}_2) - \mathbf{\Sigma}_0\right\|_F^2 &= \frac{1}{p}\left\|\text{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{P}) - \mathbf{P}_1^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{P}_1\right\|_F^2 + \frac{1}{p}\frac{\text{tr}^2[(\mathbf{T}-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}})\widetilde{\mathbf{\Sigma}}_2]}{\text{tr}(\mathbf{T}-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}})^2} \\
&\quad - \frac{2}{p}\frac{\text{tr}[(\mathbf{T}-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}})\widetilde{\mathbf{\Sigma}}_2]}{\text{tr}(\mathbf{T}-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}})^2}\text{tr}[(\mathbf{T}-\widetilde{\mathbf{\Sigma}}_2)\mathbf{\Sigma}_0] \\
&= \frac{1}{p}\sum_{i=1}^{p}(\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{p}_{1i})^2 + \frac{1}{p}\frac{\text{tr}^2[(\mathbf{T}-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}})(\widetilde{\mathbf{\Sigma}}_2-\mathbf{\Sigma}_0)]}{\text{tr}(\mathbf{T}-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}})^2} \\
&\quad - \frac{1}{p}\frac{\text{tr}^2[(\mathbf{T}-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}})\mathbf{\Sigma}_0]}{\text{tr}(\mathbf{T}-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}})^2} + \frac{1}{p}\left\|\text{diag}(\mathbf{P}_1^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{P}) - \mathbf{P}_1^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{P}_1\right\|_F^2.
\end{aligned}$$

With these, we can expand the efficiency loss as

$$\begin{aligned}
EL(\mathbf{\Sigma}_0, \mathbf{\Sigma}(\mathbf{P}_1,\mathbf{T},\widetilde{\mathbf{\Sigma}}_2)) &= 1 - \left(\frac{p^{-1}\left\|\mathbf{\Sigma}(\mathbf{P}_1,\mathbf{T},\widetilde{\mathbf{\Sigma}}_2) - \mathbf{\Sigma}_0\right\|_F^2}{p^{-1}\left\|\mathbf{\Sigma}_{\text{Ideal}} - \mathbf{\Sigma}_0\right\|_F^2}\right)^{-1} \\
&= 1 - \left(\frac{R_1+R_2+R_3+R_4}{p^{-1}\left\|\mathbf{\Sigma}_{\text{Ideal}} - \mathbf{\Sigma}_0\right\|_F^2}\right)^{-1},
\end{aligned}$$

where

$$
\begin{aligned}
R_1 &= \frac{1}{p}\left\|\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P}) - \mathbf{P}_1^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P}_1\right\|_F^2 - \frac{1}{p}\frac{\mathrm{tr}^2[(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})\boldsymbol{\Sigma}_0]}{\mathrm{tr}(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})^2},\\
R_2 &= \frac{1}{p}\sum_{i=1}^{p}(\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_{1i})^2,\\
R_3 &= \frac{1}{p}\frac{\mathrm{tr}^2[(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})\boldsymbol{\Sigma}_0]}{\mathrm{tr}(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})^2} - \frac{1}{p}\frac{\mathrm{tr}^2[(\mathbf{T}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})\boldsymbol{\Sigma}_0]}{\mathrm{tr}(\mathbf{T}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2},\\
R_4 &= \frac{1}{p}\frac{\mathrm{tr}^2[(\mathbf{T}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})(\widetilde{\boldsymbol{\Sigma}}_2-\boldsymbol{\Sigma}_0)]}{\mathrm{tr}(\mathbf{T}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2}.
\end{aligned}
$$

Firstly, we have $\frac{R_1}{p^{-1}\|\boldsymbol{\Sigma}_{\mathrm{Ideal}}-\boldsymbol{\Sigma}_0\|_F^2} \to 1$ almost surely by Lemma S.4 of Lam (2016). If we can prove that $R_2, R_3, R_4 \to 0$ in probability/almost surely, then since $p^{-1}\|\boldsymbol{\Sigma}_{\mathrm{Ideal}} - \boldsymbol{\Sigma}_0\|_F^2 \nrightarrow 0$ as $\boldsymbol{\Sigma}_0 \neq \sigma^2\mathbf{I}_p$ (see the proof of Lemma S.4 in Lam (2016) for more details), the proof will be completed.

To this end, apply Lemma 3.1 with $\mathbf{q}_i = \mathbf{p}_{1i}$,

$$
\begin{aligned}
R_2 &\le \left(\max_{1\le i\le p}\left|\frac{\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_{1i}}\right|\right)^2 \cdot \max_{1\le i\le p}(\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_{1i})^2\\
&\le \left(\max_{1\le i\le p}\left|\frac{\mathbf{p}_{1i}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_{1i}}\right|\right)^2 \cdot \|\boldsymbol{\Sigma}_0\|^2 \to 0 \qquad (3.24)
\end{aligned}
$$

almost surely as $\|\boldsymbol{\Sigma}_0\| = O(1)$ by Assumption (A2). For the term R_3, consider

$$
\begin{aligned}
R_3 &= R_{3,1} + R_{3,2}, \text{ where}\\
R_{3,1} &= \frac{1}{p}(\mathrm{tr}(\mathbf{T}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2 - \mathrm{tr}(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})^2)\cdot\frac{\mathrm{tr}^2(\mathbf{T}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})\boldsymbol{\Sigma}_0}{\mathrm{tr}(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})^2\mathrm{tr}(\mathbf{T}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2},\\
R_{3,2} &= \frac{1}{p}\frac{\mathrm{tr}^2(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})\boldsymbol{\Sigma}_0 - \mathrm{tr}^2(\mathbf{T}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})\boldsymbol{\Sigma}_0}{\mathrm{tr}(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})^2}.
\end{aligned}
$$

To bound $R_{3,1}$, consider

$$
\begin{aligned}
|R_{3,1}| &= \frac{1}{p}|\mathrm{tr}(\mathbf{T}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2 - \mathrm{tr}(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})^2| \cdot \frac{\mathrm{tr}^2[(\mathbf{T}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})\boldsymbol{\Sigma}_0]}{\mathrm{tr}(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})^2\mathrm{tr}(\mathbf{T}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2} \\
&\leq \frac{1}{p}|\mathrm{tr}(\mathbf{T}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2 - \mathrm{tr}(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})^2| \cdot \frac{p^{-1}\mathrm{tr}(\boldsymbol{\Sigma}_0^2)}{p^{-1}\mathrm{tr}(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})^2} \\
&= \left|\frac{1}{p}(\mathrm{tr}(\mathbf{T}^2) - \sum_{i=1}^{p}(\mathbf{p}_i^{\mathrm{T}}\mathbf{T}\mathbf{p}_i)^2) - \frac{1}{p}(\mathrm{tr}(\mathbf{T}^2) - \sum_{i=1}^{p}(\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{T}\mathbf{p}_{1i})^2)\right| \cdot \frac{p^{-1}\mathrm{tr}(\boldsymbol{\Sigma}_0^2)}{p^{-1}\mathrm{tr}(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})^2} \\
&\leq \max_{1\leq i\leq p}|\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{T}\mathbf{p}_{1i} + \mathbf{p}_i^{\mathrm{T}}\mathbf{T}\mathbf{p}_i| \cdot \frac{1}{p}\sum_{i=1}^{p}|\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{T}\mathbf{p}_{1i} - \mathbf{p}_i^{\mathrm{T}}\mathbf{T}\mathbf{p}_i| \cdot \frac{\|\boldsymbol{\Sigma}_0\|^2}{p^{-1}\mathrm{tr}(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})^2} \\
&\leq 2\|\mathbf{T}\| \cdot \frac{1}{p}\sum_{i=1}^{p}|\mathbf{p}_{1i}^{\mathrm{T}}\mathbf{T}\mathbf{p}_{1i} - \mathbf{p}_i^{\mathrm{T}}\mathbf{T}\mathbf{p}_i| \cdot \frac{\|\boldsymbol{\Sigma}_0\|^2}{p^{-1}\mathrm{tr}(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})^2} \to 0
\end{aligned}
$$

in probability/almost surely by the assumption $p^{-1}\mathrm{tr}(\mathbf{T}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})^2 \nrightarrow 0$ in probability/alsmost surely with $\|\mathbf{T}\|$ being finite, that $\|\boldsymbol{\Sigma}_0\| = O(1)$ by Assumption (A2), and finally by the result of Lemma 3.2.

Also, we have

$$
\begin{aligned}
&\frac{1}{p}|\mathrm{tr}[(\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})\boldsymbol{\Sigma}_0]| \\
&= \frac{1}{p}|\mathrm{tr}(\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{T}\mathbf{P})\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P}) - \mathrm{tr}(\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\mathbf{T}\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P}_1| \\
&\leq \frac{1}{p}\sum_{i=1}^{p}|\mathbf{p}_i^{\mathrm{T}}\mathbf{T}\mathbf{p}_i| \cdot |\mathbf{p}_i^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_i - \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_{1i}| + \frac{1}{p}\sum_{i=1}^{p}|\mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_{1i}| \cdot |\mathbf{p}_i^{\mathrm{T}}\mathbf{T}\mathbf{p}_i - \mathbf{p}_{1i}^{\mathrm{T}}\mathbf{T}\mathbf{p}_{1i}| \\
&\leq \|\mathbf{T}\| \cdot \frac{1}{p}\sum_{i=1}^{p}|\mathbf{p}_i^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_i - \mathbf{p}_{1i}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_{1i}| + \|\boldsymbol{\Sigma}_0\| \cdot \frac{1}{p}\sum_{i=1}^{p}|\mathbf{p}_i^{\mathrm{T}}\mathbf{T}\mathbf{p}_i - \mathbf{p}_{1i}^{\mathrm{T}}\mathbf{T}\mathbf{p}_{1i}| \to 0 \qquad (3.25)
\end{aligned}
$$

in probability/almost surely by Lemma 3.2. We can then bound $R_{3,2}$ by noting that

$$
\begin{aligned}
|R_{3,2}| &= \frac{1}{p}\frac{|\mathrm{tr}^2[(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})\boldsymbol{\Sigma}_0] - \mathrm{tr}^2[(\mathbf{T}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})\boldsymbol{\Sigma}_0]|}{\mathrm{tr}(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})^2} \\
&= \left|\frac{1}{p}\mathrm{tr}[(\mathbf{T}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})\boldsymbol{\Sigma}_0] - \frac{1}{p}\mathrm{tr}[(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})\boldsymbol{\Sigma}_0]\right| \\
&\quad \cdot \frac{|2\mathrm{tr}[(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})\boldsymbol{\Sigma}_0] + \mathrm{tr}[(\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})\boldsymbol{\Sigma}_0]|}{\mathrm{tr}(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})^2} \\
&\leq \frac{1}{p}|\mathrm{tr}[(\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})\boldsymbol{\Sigma}_0]| \cdot \left(\frac{2\|\boldsymbol{\Sigma}_0\|}{(p^{-1}\mathrm{tr}(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})^2)^{1/2}} + \frac{|p^{-1}\mathrm{tr}[(\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}}-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}})\boldsymbol{\Sigma}_0]|}{p^{-1}\mathrm{tr}(\mathbf{T}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}})^2}\right) \to 0
\end{aligned}
$$

in probability/almost surely by what we have just proved that $p^{-1}\text{tr}[(\widehat{\mathbf{\Sigma}}_{\mathbf{T}} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}})\mathbf{\Sigma}_0] \to 0$ in probability/almost surely, that $\|\mathbf{\Sigma}_0\| = O(1)$ by Assumption (A2), and the assumption that $p^{-1}\text{tr}(\mathbf{T} - \widehat{\mathbf{\Sigma}}_{\mathbf{T}})^2 \nrightarrow 0$ in probability/almost surely.

Combining, hence we have $R_3 = R_{3,1} + R_{3,2} \to 0$ in probability/almost surely.

Finally, similar to the proof of Theorem 3.2, assuming $\mathbf{T} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}} = \mathbf{Q}\mathbf{D}_{\mathbf{T}}\mathbf{Q}^{\text{T}}$, we have

$$\begin{aligned} R_4 &= \frac{1}{p}\frac{[\sum_{i=1}^p d_i(\mathbf{q}_i^{\text{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{q}_i - \mathbf{q}_i^{\text{T}}\mathbf{\Sigma}_0\mathbf{q}_i)]^2}{\text{tr}(\mathbf{T} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}})^2} \\ &\le \frac{p^{-1}\sum_{i=1}^p(\mathbf{q}_i^{\text{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{q}_i - \mathbf{q}_i^{\text{T}}\mathbf{\Sigma}_0\mathbf{q}_i)^2 \cdot p^{-1}\sum_{i=1}^p d_i^2}{p^{-1}\text{tr}(\mathbf{T} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}})^2} \\ &\le \max_{1\le i\le p}\left|\frac{\mathbf{q}_i^{\text{T}}\widetilde{\mathbf{\Sigma}}_2\mathbf{q}_i - \mathbf{q}_i^{\text{T}}\mathbf{\Sigma}_0\mathbf{q}_i}{\mathbf{q}_i^{\text{T}}\mathbf{\Sigma}_0\mathbf{q}_i}\right| \cdot p^{-1}\sum_{i=1}^p(\mathbf{q}_i^{\text{T}}\mathbf{\Sigma}_0\mathbf{q}_i)^2 \to 0 \end{aligned}$$

in probability/almost surely by Lemma 1, the fact that $p^{-1}\text{tr}(\mathbf{T} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}})^2 \nrightarrow 0$ in probability/alsmost surely by assumption, and $p^{-1}\sum_{i=1}^p(\mathbf{q}_i^{\text{T}}\mathbf{\Sigma}_0\mathbf{q}_i)^2 \le \|\mathbf{\Sigma}_0\|^2 < \infty$ by Assumption (A2). This completes the proof of the Theorem. □

Proof of Theorem 3.4. Similar to the proof of Theorem 3.1, define $R(\delta_1, \delta_2, \mathbf{D})$ as the squared Frobenius norm of $\mathbf{\Sigma}(\delta_1, \delta_2, \mathbf{D}) - \mathbf{\Sigma}_0$ in equation (3.8). Then

$$\begin{aligned} R(\delta_1, \delta_2, \mathbf{D}) = (1-\delta_1-\delta_2)^2\|\mathbf{P}\mathbf{D}\mathbf{P}^{\text{T}} - \mathbf{\Sigma}_0\|_F^2 + \delta_1^2\|\mathbf{T}_1 - \mathbf{\Sigma}_0\|_F^2 + \delta_2^2\|\mathbf{T}_2 - \mathbf{\Sigma}_0\|_F^2 \\ + 2\delta_1(1-\delta_1-\delta_2)\text{tr}[(\mathbf{P}\mathbf{D}\mathbf{P}^{\text{T}} - \mathbf{\Sigma}_0)(\mathbf{T}_1 - \mathbf{\Sigma}_0)] \\ + 2\delta_2(1-\delta_1-\delta_2)\text{tr}[(\mathbf{P}\mathbf{D}\mathbf{P}^{\text{T}} - \mathbf{\Sigma}_0)(\mathbf{T}_2 - \mathbf{\Sigma}_0)] \\ + 2\delta_1\delta_2\text{tr}[(\mathbf{T}_1 - \mathbf{\Sigma}_0)(\mathbf{T}_2 - \mathbf{\Sigma}_0)]. \end{aligned} \tag{3.26}$$

To minimize $R(\delta_1, \delta_2, \mathbf{D})$, we first calculate the solution for $\mathbf{D} = \text{diag}(d_1, \ldots, d_p)$ in terms of δ_1 and δ_2. For each $j = 1, \ldots, p$, we have

$$
\begin{aligned}
\frac{\partial R(\delta_1,\delta_2,\mathbf{D})}{\partial d_j} &= (1-\delta_1-\delta_2)^2\frac{\partial}{\partial d_j}(\mathrm{tr}(\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}}\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}})-2\mathrm{tr}(\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0))\\
&\quad+2\delta_1(1-\delta_1-\delta_2)\frac{\partial}{\partial d_j}\mathrm{tr}(\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}}(\mathbf{T}_1-\boldsymbol{\Sigma}_0))\\
&\quad+2\delta_2(1-\delta_1-\delta_2)\frac{\partial}{\partial d_j}\mathrm{tr}(\mathbf{P}\mathbf{D}\mathbf{P}^{\mathrm{T}}(\mathbf{T}_2-\boldsymbol{\Sigma}_0))\\
&= (1-\delta_1-\delta_2)^2\frac{\partial}{\partial d_j}(\sum_{i=1}^{p}d_i^2-2\sum_{i=1}^{p}d_i\mathbf{p}_i^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_i)\\
&\quad+2\delta_1(1-\delta_1-\delta_2)\frac{\partial}{\partial d_j}(\sum_{i=1}^{p}d_i\mathbf{p}_i^{\mathrm{T}}(\mathbf{T}_1-\boldsymbol{\Sigma}_0)\mathbf{p}_i)\\
&\quad+2\delta_2(1-\delta_1-\delta_2)\frac{\partial}{\partial d_j}(\sum_{i=1}^{p}d_i\mathbf{p}_i^{\mathrm{T}}(\mathbf{T}_2-\boldsymbol{\Sigma}_0)\mathbf{p}_i)\\
&= 2(1-\delta_1-\delta_2)[(1-\delta_1-\delta_2)(d_j-\mathbf{p}_j^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_j)\\
&\quad+\delta_1\mathbf{p}_j^{\mathrm{T}}(\mathbf{T}_1-\boldsymbol{\Sigma}_0)\mathbf{p}_j+\delta_2\mathbf{p}_j^{\mathrm{T}}(\mathbf{T}_2-\boldsymbol{\Sigma}_0)\mathbf{p}_j].
\end{aligned}
$$

Set $\frac{\partial R(\delta_1,\delta_2,\mathbf{D})}{\partial d_j}=0$, then since we assumed $\delta_1+\delta_2\neq 1$,

$$
\begin{aligned}
d_j &= \frac{(1-\delta_1-\delta_2)\mathbf{p}_j^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_j-\delta_1\mathbf{p}_j^{\mathrm{T}}(\mathbf{T}_1-\boldsymbol{\Sigma}_0)\mathbf{p}_j-\delta_2\mathbf{p}_j^{\mathrm{T}}(\mathbf{T}_2-\boldsymbol{\Sigma}_0)\mathbf{p}_j}{1-\delta_1-\delta_2}\\
&= \frac{1}{1-\delta_1-\delta_2}(\mathbf{p}_j^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_j-\delta_1\mathbf{p}_j^{\mathrm{T}}\mathbf{T}_1\mathbf{p}_j-\delta_2\mathbf{p}_j^{\mathrm{T}}\mathbf{T}_2\mathbf{p}_j),\ \text{so that}\\
\mathbf{D} &= \frac{1}{1-\delta_1-\delta_2}(\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})-\delta_1\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{T}_1\mathbf{P})-\delta_2\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{T}_2\mathbf{P})). \qquad (3.27)
\end{aligned}
$$

This proves the first part of Theorem 3.4. equation (3.27) can be substituted into $R(\delta_1,\delta_2,\mathbf{D})$ in (3.26) for solving for the optimal δ_1 and δ_2 then. We have

$$
\begin{aligned}
R(\delta_1,\delta_2,\mathbf{D}) &= R(\delta_1,\delta_2)\\
&= \|\mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}} - \delta_1\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_1} - \delta_2\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_2} + \delta_1\mathbf{T}_1 + \delta_2\mathbf{T}_2 - \boldsymbol{\Sigma}_0\|_F^2\\
&= \|(\mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}} - \boldsymbol{\Sigma}_0) - \delta_1(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_1} - \mathbf{T}_1) - \delta_2(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_2} - \mathbf{T}_2)\|_F^2\\
&= \mathrm{tr}(\mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}} - \boldsymbol{\Sigma}_0)^2 + \delta_1^2\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_1} - \mathbf{T}_1)^2 + \delta_2^2\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_2} - \mathbf{T}_2)^2\\
&\quad - 2\delta_1\mathrm{tr}[(\mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}} - \boldsymbol{\Sigma}_0)(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_1} - \mathbf{T}_1)]\\
&\quad - 2\delta_2\mathrm{tr}[(\mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}} - \boldsymbol{\Sigma}_0)(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_2} - \mathbf{T}_2)]\\
&\quad + 2\delta_1\delta_2\mathrm{tr}[(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_1} - \mathbf{T}_1)(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_2} - \mathbf{T}_2)], \qquad (3.28)
\end{aligned}
$$

where $\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_i} = \mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{T}_i\mathbf{P})\mathbf{P}^{\mathrm{T}}, i = 1, 2$. To find the optimal δ_1 and δ_2, we take the partial derivative of R with respect to δ_1 and δ_2 respectively,

$$
\begin{aligned}
\frac{\partial R(\delta_1,\delta_2)}{\partial\delta_1} &= 2\delta_1\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_1} - \mathbf{T}_1)^2 - 2\mathrm{tr}[(\mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}} - \boldsymbol{\Sigma}_0)(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_1} - \mathbf{T}_1)]\\
&\quad + 2\delta_2\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_1} - \mathbf{T}_1)(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_2} - \mathbf{T}_2),\\
\frac{\partial R(\delta_1,\delta_2)}{\partial\delta_2} &= 2\delta_2\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_2} - \mathbf{T}_2)^2 - 2\mathrm{tr}[(\mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}} - \boldsymbol{\Sigma}_0)(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_2} - \mathbf{T}_2)]\\
&\quad + 2\delta_1\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_1} - \mathbf{T}_1)(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_2} - \mathbf{T}_2).
\end{aligned}
$$

Setting $\frac{\partial R(\delta_1,\delta_2)}{\partial\delta_1} = 0$ and $\frac{\partial R(\delta_1,\delta_2)}{\partial\delta_2} = 0$, we get

$$
\begin{aligned}
&\delta_1\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_1} - \mathbf{T}_1)^2 + \delta_2\mathrm{tr}[(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_1} - \mathbf{T}_1)(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_2} - \mathbf{T}_2)]\\
=&\mathrm{tr}[(\mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}} - \boldsymbol{\Sigma}_0)(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_1} - \mathbf{T}_1)],\\
&\delta_2\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_2} - \mathbf{T}_2)^2 + \delta_1\mathrm{tr}[(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_1} - \mathbf{T}_1)(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_2} - \mathbf{T}_2)]\\
=&\mathrm{tr}[(\mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}} - \boldsymbol{\Sigma}_0)(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_2} - \mathbf{T}_2)].
\end{aligned}
$$

Similar to (3.20), $\mathrm{tr}[\mathbf{P}\mathrm{diag}(\mathbf{P}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\mathrm{T}}(\hat{\boldsymbol{\Sigma}}_{\mathbf{T}_i} - \mathbf{T}_i)] = 0, i = 1, 2$. Solving for δ_1 and δ_2, we get the results as stated in Theorem 3.4. The denominator in δ_1 and δ_2 is not 0 since $\mathbf{T}_1$, $\mathbf{T}_2$ and $\mathbf{T}_1 - \mathbf{T}_2$ are not of the form $\mathbf{PDP}^{\mathrm{T}}$ for some diagonal matrix $\mathbf{D}$. □

Proof of Theorem 3.5. Let Assumptions (A1) and (A2) be satisfied and the data is not from a factor model. Write $\mathbf{T}_i - \tilde{\boldsymbol{\Sigma}}_{\mathbf{T}_i} = \mathbf{Q}_i\mathbf{D}_{\mathbf{T}_i}\mathbf{Q}_i^{\mathrm{T}}$, where $\mathbf{Q}_i = (\mathbf{q}_{i1}, \ldots, \mathbf{q}_{ip})$ is orthogonal and $\mathbf{D}_{\mathbf{T}_i} = \mathrm{diag}(d_{i1}, \ldots, d_{ip})$. Then for $i, j = 1, 2$, using the notation

$$a_{ij} = \mathrm{tr}[(\mathbf{T}_i - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_i})(\mathbf{T}_j - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_j})],$$

$$\begin{aligned}
|\widehat{\delta}_i - \delta_i| &\leq \frac{|\mathrm{tr}[(\mathbf{T}_i - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_i})(\mathbf{\Sigma}_0 - \widetilde{\mathbf{\Sigma}}_2)]a_{3-i,3-i}| + |\mathrm{tr}[(\mathbf{T}_{3-i} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{3-i}})(\mathbf{\Sigma}_0 - \widetilde{\mathbf{\Sigma}}_2)]a_{12}|}{a_{22}a_{11} - a_{12}^2} \\
&\leq \frac{\left|\frac{\mathrm{tr}[(\mathbf{T}_i - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_i})(\mathbf{\Sigma}_0 - \widetilde{\mathbf{\Sigma}}_2)]}{\mathrm{tr}(\mathbf{T}_i - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_i})^2} \cdot p^{-2}a_{11}a_{22}\right| + \left|\frac{\mathrm{tr}[(\mathbf{T}_{3-i} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{3-i}})(\mathbf{\Sigma}_0 - \widetilde{\mathbf{\Sigma}}_2)]}{\mathrm{tr}(\mathbf{T}_{3-i} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{3-i}})^2} \cdot p^{-2}a_{12}a_{3-i,3-i}\right|}{p^{-2}(a_{22}a_{11} - a_{12}^2)} \\
&\to 0
\end{aligned}$$

in probability/almost surely, where we have used the same lines of proof in (3.21), and the assumption that $p^{-1}a_{ii}, p^{-2}(a_{22}a_{11} - a_{12}^2) \nrightarrow 0$ and are finite. This proves $\widehat{\delta}_i - \delta_i \to 0$ in probability/almost surely.

Now assume $\|\mathbf{T}_1 - \mathbf{\Sigma}_0\| \to 0$ in probability/almost surely. Then using $\mathrm{tr}[(\mathbf{T}_i - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_i})\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_j}] = 0$ for $i, j = 1, 2$,

$$\widehat{\delta}_1 = \frac{\mathrm{tr}[(\mathbf{T}_1 - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})(\mathbf{\Sigma}_0 - \mathbf{T}_1)]a_{22} + a_{11}a_{22} - \mathrm{tr}[(\mathbf{T}_2 - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_2})(\mathbf{\Sigma}_0 - \mathbf{T}_1)]a_{12} - a_{12}^2}{a_{22}a_{11} - a_{12}^2},$$

with

$$|\widehat{\delta}_1 - 1| \leq \frac{(p^{-1}a_{11})^{1/2} \cdot p^{-1}a_{22} \cdot \|\mathbf{T}_1 - \mathbf{\Sigma}_0\| + (p^{-1}a_{22})^{1/2} \cdot |p^{-1}a_{12}| \cdot \|\mathbf{T}_1 - \mathbf{\Sigma}_0\|}{p^{-2}(a_{22}a_{11} - a_{12}^2)} \to 0$$

in probability/almost surely, where we use $p^{-1}a_{ii}, p^{-2}(a_{22}a_{11} - a_{12}^2) \nrightarrow 0$ and are finite. At the same time,

$$\widehat{\delta}_2 = \frac{\mathrm{tr}[(\mathbf{T}_2 - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_2})(\mathbf{\Sigma}_0 - \mathbf{T}_1)]a_{11} + a_{12}a_{11} - \mathrm{tr}[(\mathbf{T}_1 - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})(\mathbf{\Sigma}_0 - \mathbf{T}_1)]a_{12} - a_{12}a_{11}}{a_{22}a_{11} - a_{12}^2},$$

with

$$|\widehat{\delta}_2| \leq \frac{(p^{-1}a_{22})^{1/2} \cdot p^{-1}a_{11} \cdot \|\mathbf{T}_1 - \mathbf{\Sigma}_0\| + (p^{-1}a_{11})^{1/2} \cdot |p^{-1}a_{12}| \cdot \|\mathbf{T}_1 - \mathbf{\Sigma}_0\|}{p^{-2}(a_{22}a_{11} - a_{12}^2)} \to 0$$

in probability/almost surely. If $\|\mathbf{T}_2 - \mathbf{\Sigma}_0\| \to 0$ in probability/almost surely, the lines to follow are exactly the same with the roles of $\widehat{\delta}_1$ and $\widehat{\delta}_2$ swapped. Hence we have proved that if $\|\mathbf{T}_i - \mathbf{\Sigma}_0\| \to 0$ in probability/almost surely, we have $\widehat{\delta}_i \to 1$ and $\widehat{\delta}_{3-i} \to 0$ in probability/almost surely.

With $\|\mathbf{T}_i - \mathbf{\Sigma}_0\| \to 0$ in probabiliy/almost surely assumed, consider

$$\begin{aligned}\|\mathbf{\Sigma}(\mathbf{P}_1, \mathbf{T}, \widetilde{\mathbf{\Sigma}}_2) - \mathbf{\Sigma}_0\| &\leq \|\mathbf{P}_1 \mathrm{diag}(\mathbf{P}_1^{\mathrm{T}} \widetilde{\mathbf{\Sigma}}_2 \mathbf{P}_1) \mathbf{P}_1^{\mathrm{T}} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_i}\| + \|\mathbf{T}_i - \mathbf{\Sigma}_0\| \\ &\quad + |\widehat{\delta}_i - 1| \cdot \|\mathbf{T}_i - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_i}\| + |\widehat{\delta}_{3-i}| \cdot \|\mathbf{T}_{3-i} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{3-i}}\| \\ &\leq \max_{1 \leq i \leq p} \left| \frac{\mathbf{p}_{1i}^{\mathrm{T}} \widetilde{\mathbf{\Sigma}}_2 \mathbf{p}_{1i} - \mathbf{p}_{1i}^{\mathrm{T}} \mathbf{\Sigma}_0 \mathbf{p}_{1i}}{\mathbf{p}_{1i}^{\mathrm{T}} \mathbf{\Sigma}_0 \mathbf{p}_{1i}} \right| \cdot \|\mathbf{\Sigma}_0\| + 2\|\mathbf{T}_i - \mathbf{\Sigma}_0\| \\ &\quad + 2|\widehat{\delta}_i - 1| \cdot \|\mathbf{T}_i\| + 2|\widehat{\delta}_{3-i}| \cdot \|\mathbf{T}_{3-i}\| \to 0\end{aligned}$$

in probability/almost surely, where we used Lemma 1 of Lam (2016) and Assumption (A2), as well as the finiteness of $\|\mathbf{T}_i\|$ in probability/almost surely.

For data from a factor model with Assumptions (F1) and (F2) in place, the proof follows exactly the same as before, except that we now have $p^{-2}a_{ii}, p^{-4}(a_{11}a_{22} - a_{12}^2) \nrightarrow 0$ while $p^{-1}\|\mathbf{\Sigma}_0\|$, $p^{-1}\|\mathbf{T}_i\|$ are finite in probability/almost surely. This completes the proof of the theorem. □

Proof of Theorem 3.6. Recall the notations $a_{ij} = \mathrm{tr}[(\mathbf{T}_i - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_i})(\mathbf{T}_j - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_j})]$ and $b_{ij} = \mathrm{tr}[(\mathbf{T}_i - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_i})(\mathbf{T}_j - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_j})]$ for $i, j = 1, 2$. In this proof, we define δ_i for $i = 1, 2$ which corresponds to the $\mathbf{\Sigma}_{\mathrm{Ideal}}$ constructed using $\mathbf{P}$ rather than $\mathbf{P}_1$, and at the same time with knowledge of $\mathbf{\Sigma}_0$ itself. Hence we have

$$\begin{aligned}\delta_i &= \frac{b_{3-i,3-i}\mathrm{tr}[(\mathbf{T}_i - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_i})\mathbf{\Sigma}_0] - b_{12}\mathrm{tr}[(\mathbf{T}_{3-i} - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_{3-i}})\mathbf{\Sigma}_0]}{b_{22}b_{11} - b_{12}^2}, \\ \hat{\delta}_i &= \frac{a_{3-i,3-i}\mathrm{tr}[(\mathbf{T}_i - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_i})\widetilde{\mathbf{\Sigma}}_2] - a_{12}\mathrm{tr}[(\mathbf{T}_{3-i} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{3-i}})\widetilde{\mathbf{\Sigma}}_2]}{a_{22}a_{11} - a_{12}^2}.\end{aligned}$$

We first consider

$$\begin{aligned}
\frac{1}{p}\left\|\boldsymbol{\Sigma}_{\text{Ideal}}-\boldsymbol{\Sigma}_0\right\|_F^2 &= \frac{1}{p}\left\|\mathbf{P}\text{diag}(\mathbf{P}^{\text{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\text{T}}-\boldsymbol{\Sigma}_0+\delta_1(\mathbf{T}_1-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_1})+\delta_2(\mathbf{T}_2-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\right\|_F^2 \\
&= \frac{1}{p}\left\|\text{diag}(\mathbf{P}^{\text{T}}\boldsymbol{\Sigma}_0\mathbf{P})-\mathbf{P}^{\text{T}}\boldsymbol{\Sigma}_0\mathbf{P}\right\|_F^2+\frac{1}{p}\left\|\delta_1(\mathbf{T}_1-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_1})+\delta_2(\mathbf{T}_2-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\right\|_F^2 \\
&\quad+\frac{2}{p}\text{tr}[(\mathbf{P}\text{diag}(\mathbf{P}^{\text{T}}\boldsymbol{\Sigma}_0\mathbf{P})\mathbf{P}^{\text{T}}-\boldsymbol{\Sigma}_0)(\delta_1(\mathbf{T}_1-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_1})+\delta_2(\mathbf{T}_2-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_2}))] \\
&= \frac{1}{p}\left\|\text{diag}(\mathbf{P}^{\text{T}}\boldsymbol{\Sigma}_0\mathbf{P})-\mathbf{P}^{\text{T}}\boldsymbol{\Sigma}_0\mathbf{P}\right\|_F^2+\frac{1}{p}(\delta_1^2b_{11}+\delta_2^2b_{22}+2\delta_1\delta_2b_{12}) \\
&\quad-\frac{2}{p}\cdot(\delta_1\text{tr}[(\mathbf{T}_1-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_1})\boldsymbol{\Sigma}_0]+\delta_2\text{tr}[(\mathbf{T}_2-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\boldsymbol{\Sigma}_0]) \\
&= \frac{1}{p}\Big(\left\|\text{diag}(\mathbf{P}^{\text{T}}\boldsymbol{\Sigma}_0\mathbf{P})-\mathbf{P}^{\text{T}}\boldsymbol{\Sigma}_0\mathbf{P}\right\|_F^2-\delta_1\text{tr}[(\mathbf{T}_1-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_1})\boldsymbol{\Sigma}_0] \\
&\quad-\delta_2\text{tr}[(\mathbf{T}_2-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\boldsymbol{\Sigma}_0]\Big),
\end{aligned}$$

where the last equality uses the result

$$\delta_1^2b_{11}+\delta_2^2b_{22}+2\delta_1\delta_2b_{12}=\delta_1\text{tr}[(\mathbf{T}_1-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_1})\boldsymbol{\Sigma}_0]+\delta_2\text{tr}[(\mathbf{T}_2-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\boldsymbol{\Sigma}_0],$$

which can be proved with simple algebra.

Similarly, we have

$$
\begin{aligned}
&\frac{1}{p}\left\|\boldsymbol{\Sigma}(\mathbf{P}_1,\mathbf{T},\widetilde{\boldsymbol{\Sigma}}_2)-\boldsymbol{\Sigma}_0\right\|_F^2\\
&=\frac{1}{p}\left\|\mathbf{P}_1\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}}-\boldsymbol{\Sigma}_0+\hat{\delta}_1(\mathbf{T}_1-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_1})+\hat{\delta}_2(\mathbf{T}_2-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\right\|_F^2\\
&=\frac{1}{p}\left\|\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{P}_1)-\mathbf{P}_1^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P}_1\right\|_F^2+\frac{1}{p}\left\|\hat{\delta}_1(\mathbf{T}_1-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_1})+\hat{\delta}_2(\mathbf{T}_2-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\right\|_F^2\\
&\quad+\frac{2}{p}\mathrm{tr}[(\mathbf{P}_1\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{P}_1)\mathbf{P}_1^{\mathrm{T}}-\boldsymbol{\Sigma}_0)(\hat{\delta}_1(\mathbf{T}_1-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_1})+\hat{\delta}_2(\mathbf{T}_2-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_2}))]\\
&=\frac{1}{p}\sum_{r=1}^{p}(\mathbf{p}_{1r}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1r}-\mathbf{p}_{1r}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_{1r})^2+\frac{1}{p}\left\|\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P}_1)-\mathbf{P}_1^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P}_1\right\|_F^2\\
&\quad+\frac{1}{p}a_{22}\mathrm{tr}[(\mathbf{T}_1-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_1})\widetilde{\boldsymbol{\Sigma}}_2]\cdot\frac{\mathrm{tr}[(\mathbf{T}_1-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_1})\widetilde{\boldsymbol{\Sigma}}_2]-2\mathrm{tr}[(\mathbf{T}_1-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_1})\boldsymbol{\Sigma}_0]}{a_{22}a_{11}-a_{12}^2}\\
&\quad+\frac{1}{p}a_{11}\mathrm{tr}[(\mathbf{T}_2-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\widetilde{\boldsymbol{\Sigma}}_2]\cdot\frac{\mathrm{tr}[(\mathbf{T}_2-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\widetilde{\boldsymbol{\Sigma}}_2]-2\mathrm{tr}[(\mathbf{T}_2-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\boldsymbol{\Sigma}_0]}{a_{22}a_{11}-a_{12}^2}\\
&\quad+\frac{2}{p}a_{12}\Bigg(\frac{\mathrm{tr}[(\mathbf{T}_2-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\widetilde{\boldsymbol{\Sigma}}_2]\mathrm{tr}[(\mathbf{T}_1-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_1})\boldsymbol{\Sigma}_0]}{a_{22}a_{11}-a_{12}^2}\\
&\quad+\frac{\mathrm{tr}[(\mathbf{T}_1-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_1})\widetilde{\boldsymbol{\Sigma}}_2]\mathrm{tr}[(\mathbf{T}_2-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\boldsymbol{\Sigma}_0]}{a_{22}a_{11}-a_{12}^2}\\
&\quad-\frac{\mathrm{tr}[(\mathbf{T}_1-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_1})\widetilde{\boldsymbol{\Sigma}}_2]\mathrm{tr}[(\mathbf{T}_2-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\widetilde{\boldsymbol{\Sigma}}_2]}{a_{22}a_{11}-a_{12}^2}\Bigg)\\
&=R_1+R_2+R_3+R_4,\quad\text{where}
\end{aligned}
$$

$$
\begin{aligned}
R_1&=\frac{1}{p}\left(\left\|\mathrm{diag}(\mathbf{P}_1^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P})-\mathbf{P}_1^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{P}_1\right\|_F^2-\delta_1\mathrm{tr}[(\mathbf{T}_1-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_1})\boldsymbol{\Sigma}_0]-\delta_2\mathrm{tr}[(\mathbf{T}_2-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\boldsymbol{\Sigma}_0]\right),\\
R_2&=\frac{1}{p}\sum_{r=1}^{p}(\mathbf{p}_{1r}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_2\mathbf{p}_{1r}-\mathbf{p}_{1r}^{\mathrm{T}}\boldsymbol{\Sigma}_0\mathbf{p}_{1r})^2,\\
R_3&=\frac{1}{p}\Bigg((a_{22}\mathrm{tr}[(\mathbf{T}_1-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_1})\widetilde{\boldsymbol{\Sigma}}_2]-2a_{12}\mathrm{tr}[(\mathbf{T}_2-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\widetilde{\boldsymbol{\Sigma}}_2])\frac{\mathrm{tr}[(\mathbf{T}_1-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_1})(\widetilde{\boldsymbol{\Sigma}}_2-\boldsymbol{\Sigma}_0)]}{a_{22}a_{11}-a_{12}^2}\\
&\quad+(a_{11}\mathrm{tr}[(\mathbf{T}_2-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\widetilde{\boldsymbol{\Sigma}}_2]-2a_{12}\mathrm{tr}[(\mathbf{T}_1-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_1})\widetilde{\boldsymbol{\Sigma}}_2])\frac{\mathrm{tr}[(\mathbf{T}_2-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_2})(\widetilde{\boldsymbol{\Sigma}}_2-\boldsymbol{\Sigma}_0)]}{a_{22}a_{11}-a_{12}^2}\Bigg),\\
R_4&=\frac{1}{p}\Bigg(\frac{2a_{12}\mathrm{tr}[(\mathbf{T}_1-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_1})\widetilde{\boldsymbol{\Sigma}}_2]\mathrm{tr}[(\mathbf{T}_2-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\widetilde{\boldsymbol{\Sigma}}_2]}{a_{22}a_{11}-a_{12}^2}\\
&\quad-\frac{a_{22}\mathrm{tr}[(\mathbf{T}_1-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_1})\widetilde{\boldsymbol{\Sigma}}_2]\mathrm{tr}[(\mathbf{T}_1-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_1})\boldsymbol{\Sigma}_0]}{a_{22}a_{11}-a_{12}^2}\\
&\quad-\frac{a_{11}\mathrm{tr}[(\mathbf{T}_2-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\widetilde{\boldsymbol{\Sigma}}_2]\mathrm{tr}[(\mathbf{T}_2-\widetilde{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\boldsymbol{\Sigma}_0]}{a_{22}a_{11}-a_{12}^2}\\
&\quad+\delta_1\mathrm{tr}[(\mathbf{T}_1-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_1})\boldsymbol{\Sigma}_0]+\delta_2\mathrm{tr}[(\mathbf{T}_2-\widehat{\boldsymbol{\Sigma}}_{\mathbf{T}_2})\boldsymbol{\Sigma}_0]\Bigg).
\end{aligned}
$$

With these, we can expand the efficiency loss as

$$\begin{aligned} EL(\mathbf{\Sigma}_0, \mathbf{\Sigma}(\mathbf{P}_1, \mathbf{T}, \widetilde{\mathbf{\Sigma}}_2)) &= 1 - \left(\frac{p^{-1} \left\| \mathbf{\Sigma}(\mathbf{P}_1, \mathbf{T}, \widetilde{\mathbf{\Sigma}}_2) - \mathbf{\Sigma}_0 \right\|_F^2}{p^{-1} \left\| \mathbf{\Sigma}_{\text{Ideal}} - \mathbf{\Sigma}_0 \right\|_F^2} \right)^{-1} \\ &= 1 - \left(\frac{R_1 + R_2 + R_3 + R_4}{p^{-1} \left\| \mathbf{\Sigma}_{\text{Ideal}} - \mathbf{\Sigma}_0 \right\|_F^2} \right)^{-1}. \end{aligned}$$

Similar to the proof of Theorem 3.3, we have $\frac{R_1}{p^{-1}\|\mathbf{\Sigma}_{\text{Ideal}} - \mathbf{\Sigma}_0\|_F^2} \to 1$ almost surely by Lemma S.4 of Lam (2016), and $R_2 \to 0$ almost surely by Lemma 3.1, following the exact lines of proof in (3.24). The proof completes if we can prove that $R_3, R_4 \to 0$ in probability/almost surely.

To do so, we prove several intermediate results first. Firstly,

$$\begin{aligned} \frac{1}{p}|a_{ij} - b_{ij}| &= \frac{1}{p}\left|\text{tr}[(\mathbf{T}_i - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_i})(\mathbf{T}_j - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_j})] - \text{tr}[(\mathbf{T}_i - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_i})(\mathbf{T}_j - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_j})]\right| \\ &= \frac{1}{p}|\text{tr}[(\widehat{\mathbf{\Sigma}}_{\mathbf{T}_i} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_i})\mathbf{T}_j]| \\ &\le \frac{\|\mathbf{T}_i\|}{p} \sum_{r=1}^{p} |\mathbf{p}_r^{\mathrm{T}} \mathbf{T}_j \mathbf{p}_r - \mathbf{p}_{1r}^{\mathrm{T}} \mathbf{T}_j \mathbf{p}_{1r}| + \frac{\|\mathbf{T}_j\|}{p} \sum_{r=1}^{p} |\mathbf{p}_r^{\mathrm{T}} \mathbf{T}_i \mathbf{p}_r - \mathbf{p}_{1r}^{\mathrm{T}} \mathbf{T}_i \mathbf{p}_{1r}| \\ &\to 0 \end{aligned} \tag{3.29}$$

almost surely for $i, j = 1, 2$ by Lemma 3.2, and the assumption that $\|\mathbf{T}_i\|$ is finite in probability/almost surely. Secondly,

$$\frac{1}{p}|\text{tr}[(\mathbf{T}_i - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_i})(\mathbf{\Sigma}_0 - \widetilde{\mathbf{\Sigma}}_2)]| = \frac{1}{p} a_{ii} \left| \frac{\text{tr}[(\mathbf{T}_i - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_i})(\mathbf{\Sigma}_0 - \widetilde{\mathbf{\Sigma}}_2)]}{\text{tr}(\mathbf{T}_i - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_i})^2} \right| \to 0 \tag{3.30}$$

in probability/almost surely for $i = 1, 2$ by (3.21), and the assumption that $p^{-1}a_{ii}$ is finite in probability/almost surely. Also, in probability/almost surely,

$$\frac{1}{p}|\text{tr}[(\mathbf{T}_i - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_i})\mathbf{\Sigma}_0]| \le \left(\frac{1}{p}\text{tr}(\mathbf{T}_i - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_i})^2\right)^{1/2} \cdot \left(\frac{1}{p}\text{tr}(\mathbf{\Sigma}_0^2)\right)^{1/2} \le (p^{-1}b_{ii})^{1/2} \cdot \|\mathbf{\Sigma}_0\| < \infty \tag{3.31}$$

by the assumptions that $p^{-1}b_{ii}$ and $\|\mathbf{\Sigma}_0\|$ are both finite. Similarly, we have $\frac{1}{p}|\text{tr}[(\mathbf{T}_i - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_i})\mathbf{\Sigma}_0]| < \infty$ in probability/almost surely. Combining these results with (3.30), we

get

$$\frac{1}{p}|\text{tr}[(\mathbf{T}_i - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_i})\widetilde{\mathbf{\Sigma}}_2]| \leq \frac{1}{p}|\text{tr}[(\mathbf{T}_i - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_i})(\mathbf{\Sigma}_0 - \widetilde{\mathbf{\Sigma}}_2)]| + \frac{1}{p}|\text{tr}[(\mathbf{T}_i - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_i})\mathbf{\Sigma}_0]| < \infty \quad (3.32)$$

in probability/almost surely. At the same time,

$$\begin{aligned} & p^{-2}|\text{tr}^2[(\mathbf{T}_i - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_i})\mathbf{\Sigma}_0] - \text{tr}^2[(\mathbf{T}_i - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_i})\mathbf{\Sigma}_0]| \\ \leq\ & |p^{-1}\text{tr}[(\widehat{\mathbf{\Sigma}}_{\mathbf{T}} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}})\mathbf{\Sigma}_0] + 2p^{-1}\text{tr}[(\mathbf{T}_i - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_i})\mathbf{\Sigma}_0]| \cdot p^{-1}|\text{tr}[(\widehat{\mathbf{\Sigma}}_{\mathbf{T}} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}})\mathbf{\Sigma}_0]| \\ \to\ & 0 \end{aligned} \quad (3.33)$$

in probability/almost surely by (3.25) and (3.31).

We can then bound R_3 by noting that

$$\begin{aligned} |R_3| \leq\ & \left(\frac{1}{p}a_{22} \cdot \frac{1}{p}|\text{tr}[(\mathbf{T}_1 - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})\widetilde{\mathbf{\Sigma}}_2]| + \frac{2}{p}|a_{12}| \cdot \frac{1}{p}|\text{tr}[(\mathbf{T}_2 - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_2})\widetilde{\mathbf{\Sigma}}_2]|\right) \\ & \cdot \frac{p^{-1}|\text{tr}[(\mathbf{T}_1 - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})(\widetilde{\mathbf{\Sigma}}_2 - \mathbf{\Sigma}_0)]|}{p^{-2}|a_{22}a_{11} - a_{12}^2|} \\ & + \left(\frac{1}{p}a_{11} \cdot \frac{1}{p}|\text{tr}[(\mathbf{T}_2 - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_2})\widetilde{\mathbf{\Sigma}}_2]| + \frac{2}{p}|a_{12}| \cdot \frac{1}{p}|\text{tr}[(\mathbf{T}_1 - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})\widetilde{\mathbf{\Sigma}}_2]|\right) \\ & \cdot \frac{p^{-1}|\text{tr}[(\mathbf{T}_2 - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_2})(\widetilde{\mathbf{\Sigma}}_2 - \mathbf{\Sigma}_0)]|}{p^{-2}|a_{22}a_{11} - a_{12}^2|} \\ \to\ & 0 \end{aligned}$$

in probability/almost surely by the results in (3.30) and (3.32), along with the assumptions for the finiteness of $p^{-1}a_{ij}$ and $p^{-2}(a_{22}a_{11} - a_{12}^2)$ in probability/almost surely for $i, j = 1, 2$.

Also, we can decompose R_4 further. Consider

$$
\begin{aligned}
R_4 = \frac{1}{p}\Bigg(& \frac{2a_{12}\mathrm{tr}[(\mathbf{T}_1-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})\widetilde{\mathbf{\Sigma}}_2]\mathrm{tr}[(\mathbf{T}_2-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_2})\widetilde{\mathbf{\Sigma}}_2]}{a_{22}a_{11}-a_{12}^2} \\
& - \frac{a_{22}\mathrm{tr}[(\mathbf{T}_1-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})\widetilde{\mathbf{\Sigma}}_2]\mathrm{tr}[(\mathbf{T}_1-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]}{a_{22}a_{11}-a_{12}^2} \\
& - \frac{a_{11}\mathrm{tr}[(\mathbf{T}_2-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_2})\widetilde{\mathbf{\Sigma}}_2]\mathrm{tr}[(\mathbf{T}_2-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_2})\mathbf{\Sigma}_0]}{a_{22}a_{11}-a_{12}^2} \\
& - \frac{2b_{12}\mathrm{tr}[(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]\mathrm{tr}[(\mathbf{T}_2-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})\mathbf{\Sigma}_0]}{b_{22}b_{11}-b_{12}^2} \\
& + \frac{b_{22}\mathrm{tr}^2[(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]+b_{11}\mathrm{tr}^2[(\mathbf{T}_2-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})\mathbf{\Sigma}_0]}{b_{22}b_{11}-b_{12}^2}\Bigg) \\
= & \frac{2p^{-3}a_{12}\mathrm{tr}[(\mathbf{T}_1-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})\widetilde{\mathbf{\Sigma}}_2]\mathrm{tr}[(\mathbf{T}_2-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_2})\widetilde{\mathbf{\Sigma}}_2]}{p^{-2}(a_{22}a_{11}-a_{12}^2)} \\
& - \frac{p^{-3}a_{22}\mathrm{tr}[(\mathbf{T}_1-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})\widetilde{\mathbf{\Sigma}}_2]\mathrm{tr}[(\mathbf{T}_1-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]}{p^{-2}(a_{22}a_{11}-a_{12}^2)} \\
& - \frac{p^{-3}a_{11}\mathrm{tr}[(\mathbf{T}_2-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_2})\widetilde{\mathbf{\Sigma}}_2]\mathrm{tr}[(\mathbf{T}_2-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_2})\mathbf{\Sigma}_0]}{p^{-2}(a_{22}a_{11}-a_{12}^2)} \\
& - \frac{2p^{-3}b_{12}\mathrm{tr}[(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]\mathrm{tr}[(\mathbf{T}_2-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})\mathbf{\Sigma}_0]}{p^{-2}(a_{22}a_{11}-a_{12}^2)} \\
& + \frac{p^{-3}(b_{22}\mathrm{tr}^2[(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]+b_{11}\mathrm{tr}^2[(\mathbf{T}_2-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})\mathbf{\Sigma}_0])}{p^{-2}(a_{22}a_{11}-a_{12}^2)} \\
& - \frac{p^{-1}(\delta_1\mathrm{tr}[(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]+\delta_2\mathrm{tr}[(\mathbf{T}_2-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})\mathbf{\Sigma}_0])}{p^{-2}(a_{22}a_{11}-a_{12}^2)} \\
& \cdot \frac{p^{-2}(b_{22}b_{11}-b_{12}^2-a_{22}a_{11}+a_{12}^2)}{p^{-2}(a_{22}a_{11}-a_{12}^2)} \\
= & \frac{R_{4,1}+R_{4,2}+2R_{4,3}}{p^{-2}(a_{22}a_{11}-a_{12}^2)} + \frac{R_{4,4}\cdot R_{4,5}}{p^{-2}(a_{22}a_{11}-a_{12}^2)}, \quad \text{where} \\
R_{4,1} = & \, p^{-3}b_{22}\mathrm{tr}^2[(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0] - p^{-3}a_{22}\mathrm{tr}[(\mathbf{T}_1-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})\widetilde{\mathbf{\Sigma}}_2]\mathrm{tr}[(\mathbf{T}_1-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0], \\
R_{4,2} = & \, p^{-3}b_{11}\mathrm{tr}^2[(\mathbf{T}_2-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})\mathbf{\Sigma}_0] - p^{-3}a_{11}\mathrm{tr}[(\mathbf{T}_2-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_2})\widetilde{\mathbf{\Sigma}}_2]\mathrm{tr}[(\mathbf{T}_2-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_2})\mathbf{\Sigma}_0], \\
R_{4,3} = & \, p^{-3}a_{12}\mathrm{tr}[(\mathbf{T}_1-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})\widetilde{\mathbf{\Sigma}}_2]\mathrm{tr}[(\mathbf{T}_2-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_2})\widetilde{\mathbf{\Sigma}}_2] \\
& - p^{-3}b_{12}\mathrm{tr}[(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]\mathrm{tr}[(\mathbf{T}_2-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})\mathbf{\Sigma}_0], \\
R_{4,4} = & \, p^{-2}(b_{22}b_{11}-b_{12}^2-a_{22}a_{11}+a_{12}^2), \\
R_{4,5} = & -p^{-1}(\delta_1\mathrm{tr}[(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]+\delta_2\mathrm{tr}[(\mathbf{T}_2-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})\mathbf{\Sigma}_0]).
\end{aligned}
$$

To bound $R_{4,1}$, consider

$$\begin{aligned}
|R_{4,1}| &= |p^{-3}(b_{22}-a_{22})\mathrm{tr}^2[(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0] \\
&\quad + p^{-3}a_{22}(\mathrm{tr}^2[(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]-\mathrm{tr}^2[(\mathbf{T}_1-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]) \\
&\quad + p^{-3}a_{22}\mathrm{tr}[(\mathbf{T}_1-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]\mathrm{tr}[(\mathbf{T}_1-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})(\mathbf{\Sigma}_0-\widetilde{\mathbf{\Sigma}}_2)]| \\
&\le p^{-1}|b_{22}-a_{22}|\cdot p^{-2}\mathrm{tr}^2[(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0] \\
&\quad + p^{-1}a_{22}\cdot p^{-2}|\mathrm{tr}^2[(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]-\mathrm{tr}^2[(\mathbf{T}_1-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]| \\
&\quad + p^{-1}a_{22}\cdot p^{-1}|\mathrm{tr}[(\mathbf{T}_1-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]|\cdot p^{-1}|\mathrm{tr}[(\mathbf{T}_1-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})(\mathbf{\Sigma}_0-\widetilde{\mathbf{\Sigma}}_2)]| \\
&\to 0
\end{aligned}$$

in probability/almost surely by the results of (3.29), (3.30), (3.31), (3.32) and (3.33), together with the assumption that $p^{-1}a_{22}$ is finite. Follow the exactly the same proof, we can easily get $R_{4,2}\to 0$ in probability/almost surely. Furthermore,

$$\begin{aligned}
|R_{4,3}| &\le p^{-1}|b_{12}|\cdot p^{-1}|\mathrm{tr}[(\mathbf{T}_2-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_2})\widetilde{\mathbf{\Sigma}}_2]| \\
&\qquad \cdot (p^{-1}|\mathrm{tr}[(\mathbf{T}_1-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})(\widetilde{\mathbf{\Sigma}}_2-\mathbf{\Sigma}_0)]|+p^{-1}|\mathrm{tr}[(\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1}-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]|) \\
&\quad + p^{-1}|b_{12}|\cdot p^{-1}|\mathrm{tr}[(\mathbf{T}_1-\widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0]| \\
&\qquad \cdot (p^{-1}|\mathrm{tr}[(\mathbf{T}_2-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_2})(\widetilde{\mathbf{\Sigma}}_2-\mathbf{\Sigma}_0)]|+p^{-1}|\mathrm{tr}[(\widehat{\mathbf{\Sigma}}_{\mathbf{T}_2}-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_2})\mathbf{\Sigma}_0]|) \\
&\quad + p^{-1}|a_{12}-b_{12}|\cdot p^{-1}|\mathrm{tr}[(\mathbf{T}_1-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_1})\widetilde{\mathbf{\Sigma}}_2]|\cdot p^{-1}|\mathrm{tr}[(\mathbf{T}_2-\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_2})\widetilde{\mathbf{\Sigma}}_2]| \\
&\to 0
\end{aligned}$$

in probability/almost surely by the results of (3.25), (3.29), (3.30), (3.31) and (3.32), together with the assumption that $p^{-1}b_{12}$ is finite in probability/almost surely.

$$\begin{aligned}
|R_{4,4}| &= p^{-2}(b_{22}b_{11}-b_{12}^2-a_{22}a_{11}+a_{12}^2) \\
&= p^{-2}(b_{11}(b_{22}-a_{22})+a_{22}(b_{11}-a_{11})+(a_{12}+b_{12})(a_{12}-b_{12})) \\
&\le p^{-1}b_{11}\cdot p^{-1}|b_{22}-a_{22}|+p^{-1}a_{22}\cdot p^{-1}|b_{11}-a_{11}|+p^{-1}|b_{12}+a_{12}|\cdot p^{-1}|b_{12}-a_{12}| \\
&\to 0
\end{aligned}$$

almost surely by the result of (3.29) and the assumption that $p^{-1}a_{ij}, p^{-1}b_{ij}$ are all finite in probability/almost surely. Also, in probability/almost surely,

$$R_{4,5} = \frac{2p^{-1}b_{12} \cdot p^{-1}\mathrm{tr}[(\mathbf{T}_1 - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0] \cdot p^{-1}\mathrm{tr}[(\mathbf{T}_2 - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})\mathbf{\Sigma}_0]}{p^{-2}(b_{22}b_{11} - b_{12}^2)} - \frac{p^{-1}b_{22} \cdot p^{-2}\mathrm{tr}^2[(\mathbf{T}_1 - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_1})\mathbf{\Sigma}_0] + p^{-1}b_{11} \cdot p^{-2}\mathrm{tr}^2[(\mathbf{T}_2 - \widehat{\mathbf{\Sigma}}_{\mathbf{T}_2})\mathbf{\Sigma}_0]}{p^{-2}(b_{22}b_{11} - b_{12}^2)} < \infty$$

by the result of (3.31) and the assumption that $p^{-1}b_{ij}$ and $p^{-2}(b_{22}b_{11} - b_{12}^2)$ are all finite in probability/almost surely for $i, j = 1, 2$.

So, $R_{4,5}$ is finite. By the assumption for the finiteness of $p^{-2}(a_{22}a_{11} - a_{12}^2)$ in probability/almost surely, we can finally conclude that $R_4 \to 0$ in probability/almost surely. This completes the proof of the theorem. □

Proof of Theorem 3.7. For Frobenius loss,

$$\left\|\hat{\mathbf{\Sigma}}_{m,M} - \mathbf{\Sigma}_0\right\|_F^2 = \left\|\frac{1}{M}\sum_{j=1}^{M}(\hat{\mathbf{\Sigma}}_m^{(j)} - \mathbf{\Sigma}_0)\right\|_F^2 \leq (\frac{1}{M}\sum_{j=1}^{M}\left\|\hat{\mathbf{\Sigma}}_m^{(j)} - \mathbf{\Sigma}_0\right\|_F)^2 \leq \frac{1}{M}\sum_{j=1}^{M}\left\|\hat{\mathbf{\Sigma}}_m^{(j)} - \mathbf{\Sigma}_0\right\|_F^2, \tag{3.34}$$

so that,

$$EL(\mathbf{\Sigma}_0, \hat{\mathbf{\Sigma}}_{m,M}) \leq 1 - \frac{\left\|\hat{\mathbf{\Sigma}}_{\text{Ideal}} - \mathbf{\Sigma}_0\right\|_F^2}{\frac{1}{M}\sum_{j=1}^{M}\left\|\hat{\mathbf{\Sigma}}_m^{(j)} - \mathbf{\Sigma}_0\right\|_F^2} \leq 1 - \frac{1}{\frac{1}{M}\sum_{j=1}^{M}\frac{1}{1 - EL(\mathbf{\Sigma}_0, \hat{\mathbf{\Sigma}}_m^{(j)})}} \to 0. \tag{3.35}$$

The last step follows Theorem 3.6 that, if $p^{-1}a_{11}, p^{-1}a_{22}$ and $p^{-2}(a_{11}a_{22} - a_{12}^2) \not\to 0$ in probability/almost surely, then $EL(\mathbf{\Sigma}_0, \hat{\mathbf{\Sigma}}_m^{(j)}) \to 0$ in probability/almost surely. This proves the first part of the theorem.

To prove the remaining part of theorem, for $i = 1, 2$ and $j = 1, \ldots, M$, we note first that

$$\mathrm{tr}(\widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{ij}}) = \mathrm{tr}(\mathbf{P}_1^{\mathrm{T}}\mathrm{diag}(\mathbf{P}_{1j}\mathbf{T}_{ij}\mathbf{P}_{1j}^{\mathrm{T}})\mathbf{P}_{1j}) = \mathrm{tr}(\mathrm{diag}(\mathbf{P}_{1j}\mathbf{T}_{ij}\mathbf{P}_{1j}^{\mathrm{T}})) = \mathrm{tr}(\mathbf{P}_{1j}\mathbf{T}_{ij}\mathbf{P}_{1j}^{\mathrm{T}}) = \mathrm{tr}(\mathbf{T}_{ij}). \tag{3.36}$$

Using (3.36), with $\widehat{\delta}_{ij}$ being finite in probability/almost surely since the corresponding $p^{-1}a_{11}, p^{-1}a_{22}$ and $p^{-2}(a_{11}a_{22} - a_{12}^2) \not\to 0$ in probability/almost surely,

$$\begin{aligned}\frac{1}{p}[\mathrm{tr}(\widehat{\mathbf{\Sigma}}_{m,M}) - \mathrm{tr}(\mathbf{\Sigma}_0)] &= \frac{1}{p}\Big[\mathrm{tr}\Big(\frac{1}{M}\sum_{j=1}^{M}\mathbf{P}_{1j}\mathrm{diag}(\mathbf{P}_{1j}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{P}_{1j})\mathbf{P}_{1j}^{\mathrm{T}}\Big) - \mathrm{tr}(\mathbf{\Sigma}_0)\Big] \\ &= \frac{1}{pM}\sum_{j=1}^{M}[\mathrm{tr}(\mathbf{P}_{1j}\mathrm{diag}(\mathbf{P}_{1j}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{P}_{1j})\mathbf{P}_{1j}^{\mathrm{T}}) - \mathrm{tr}(\mathbf{\Sigma}_0)] \\ &= \frac{1}{M}\sum_{j=1}^{M}\frac{1}{p}[\mathrm{tr}(\mathbf{P}_{1j}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{P}_{1j}) - \mathrm{tr}(\mathbf{P}_{1j}^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{P}_{1j})] \to 0, \qquad (3.37)\end{aligned}$$

where the last convergence step follows from Theorem 1 of Lam (2016) if Assumptions (A1), (A2) are satisfied, and it follows from Theorem 3 of Lam (2016) if the data follows a factor model with Assumptions (F1), (F2) satisfied.

Finally, to prove the last inequality, we first consider data under Assumptions (A1) and (A2) first. Observe that

$$\begin{aligned}\frac{1}{p}\mathrm{tr}(\widehat{\mathbf{\Sigma}}_{m,M}\mathbf{\Sigma}_0) = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{p}\Big(&\mathrm{tr}(\mathbf{P}_{1j}\mathrm{diag}(\mathbf{P}_{1j}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{P}_{1j})\mathbf{P}_{1j}^{\mathrm{T}}\mathbf{\Sigma}_0) + \hat{\delta}_{1j}\mathrm{tr}[(\mathbf{T}_{1j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{1j}})\mathbf{\Sigma}_0] \\ &+ \hat{\delta}_{2j}\mathrm{tr}[(\mathbf{T}_{2j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{2j}})\mathbf{\Sigma}_0]\Big) = \frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{5}R_{ij}, \quad \text{where}\end{aligned}$$

$$\begin{aligned}R_{1j} &= \frac{1}{p}\mathrm{tr}(\mathbf{P}_{1j}\mathrm{diag}(\mathbf{P}_{1j}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{P}_{1j})\mathbf{P}_{1j}^{\mathrm{T}}\mathbf{\Sigma}_0), \\ R_{2j} &= \frac{1}{p}\delta_{1j}\mathrm{tr}[(\mathbf{T}_{1j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{1j}})\mathbf{\Sigma}_0],\ R_{3j} = \frac{1}{p}\delta_{2j}\mathrm{tr}[(\mathbf{T}_{2j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{2j}})\mathbf{\Sigma}_0], \\ R_{4j} &= \frac{1}{p}(\hat{\delta}_{1j} - \delta_{1j})\mathrm{tr}[(\mathbf{T}_{1j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{1j}})\mathbf{\Sigma}_0],\ R_{5j} = \frac{1}{p}(\hat{\delta}_{2j} - \delta_{2j})\mathrm{tr}[(\mathbf{T}_{2j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{2j}})\mathbf{\Sigma}_0].\end{aligned}$$

We analyze each term above. Firstly, we have for $\mathbf{P}_{1j} = (\mathbf{p}_{1j,1}, \ldots, \mathbf{p}_{1j,p})$,

$$\begin{aligned}R_{1j} &= R_{1j,1} + R_{1j,2}, \quad \text{where} \\ R_{1j,1} &= \frac{1}{p}\sum_{r=1}^{p}\left(\frac{\mathbf{p}_{1j,r}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{p}_{1j,r}}{\mathbf{p}_{1j,r}^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{p}_{1j,r}} - 1\right)(\mathbf{p}_{1j,r}^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{p}_{1j,r})^2,\ R_{1j,2} = \frac{1}{p}\sum_{r=1}^{p}(\mathbf{p}_{1j,r}^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{p}_{1j,r})^2.\end{aligned}$$

The first term has

$$\begin{aligned}|R_{1j,1}| &\le \max_{1\le r\le p}\left|\frac{\mathbf{p}_{1j,r}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{p}_{1j,r}}{\mathbf{p}_{1j,r}^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{p}_{1j,r}} - 1\right| \cdot \frac{1}{p}\sum_{r=1}^{p}(\mathbf{p}_{1j,r}^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{p}_{1j,r})^2 \\ &\le \max_{1\le r\le p}\left|\frac{\mathbf{p}_{1j,r}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_2^{(j)}\mathbf{p}_{1j,r}}{\mathbf{p}_{1j,r}^{\mathrm{T}}\mathbf{\Sigma}_0\mathbf{p}_{1j,r}} - 1\right| \cdot \lambda_{\max}^2(\mathbf{\Sigma}_0) \to 0,\end{aligned}$$

where the last step used the almost sure convergence result in Lemma 1 of Lam (2016), and that $\|\mathbf{\Sigma}_0\| = O(1)$. We also have $R_{1j,2} \geq \lambda^2_{\min}(\mathbf{\Sigma}_0)$, so that we have proved almost surely,

$$R_{1j} \geq \lambda^2_{\min}(\mathbf{\Sigma}_0). \tag{3.38}$$

Next, consider

$$\begin{aligned}
&p(R_{2j} + R_{3j}) \\
&= \frac{\mathrm{tr}^2[(\mathbf{T}_{1j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{1j}})\mathbf{\Sigma}_0]\mathrm{tr}(\mathbf{T}_{2j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{2j}})^2 + \mathrm{tr}^2[(\mathbf{T}_{2j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{2j}})\mathbf{\Sigma}_0]\mathrm{tr}(\mathbf{T}_{1j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{1j}})^2}{\mathrm{tr}(\mathbf{T}_{1j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{1j}})^2\mathrm{tr}(\mathbf{T}_{2j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{2j}})^2 - \mathrm{tr}^2[(\mathbf{T}_{1j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{1j}})(\mathbf{T}_{2j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{2j}})]} \\
&\quad - \frac{2\mathrm{tr}[(\mathbf{T}_{1j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{1j}})\mathbf{\Sigma}_0]\mathrm{tr}[(\mathbf{T}_{2j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{2j}})\mathbf{\Sigma}_0]\mathrm{tr}[(\mathbf{T}_{1j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{1j}})(\mathbf{T}_{2j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{2j}})]}{\mathrm{tr}(\mathbf{T}_{1j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{1j}})^2\mathrm{tr}(\mathbf{T}_{2j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{2j}})^2 - \mathrm{tr}^2[(\mathbf{T}_{1j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{1j}})(\mathbf{T}_{2j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{2j}})]} \\
&= \frac{\mathrm{tr}\Big((\mathbf{T}_{1j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{1j}})\mathrm{tr}[(\mathbf{T}_{2j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{2j}})\mathbf{\Sigma}_0] - (\mathbf{T}_{2j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{2j}})\mathrm{tr}[(\mathbf{T}_{1j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{1j}})\mathbf{\Sigma}_0]\Big)^2}{\mathrm{tr}(\mathbf{T}_{1j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{1j}})^2\mathrm{tr}(\mathbf{T}_{2j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{2j}})^2 - \mathrm{tr}^2[(\mathbf{T}_{1j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{1j}})(\mathbf{T}_{2j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{2j}})]} \geq 0
\end{aligned}$$

in probability/almost surely, since the denominator above is also non-negative by the Cauchy-Schwarz inequality on the inner product $\mathrm{tr}(\mathbf{AB})$ where $\mathbf{A}, \mathbf{B}$ are real square matrices, and by our assumptions it is positive in probability/almost surely.

Finally,

$$|R_{4j}| \leq |\widehat{\delta}_{1j} - \delta_{1j}|\Big(\frac{1}{p}\mathrm{tr}(\mathbf{T}_{1j} - \widetilde{\mathbf{\Sigma}}_{\mathbf{T}_{1j}})^2\Big)^{1/2}\Big(\frac{1}{p}\mathrm{tr}(\mathbf{\Sigma}_0^2)\Big)^{1/2} \to 0$$

in probability/almost surely, which is the result of Theorem 3.5 and our assumptions that $p^{-1}a_{11} < \infty$ and $\|\mathbf{\Sigma}_0\| = O(1)$. Similar result holds for $|R_{5j}|$. Hence combining (3.38) and the results we proved for R_{2j} to R_{5j}, the result follows.

It remains to show the trace property for data from a factor model under Assumptions (F1) and (F2). But in fact all the above steps follow, except we need to note that now $p^{-2}a_{ii}$ and $p^{-2}\lambda_{\max}(\mathbf{\Sigma}_0)$ are finite. This completes the proof of the theorem. □

Chapter 4

A Nonparametric Eigenvalue-Regularized Integrated Covariance Matrix Estimator for Asset Return Data

Declaration This chapter is based on joint work with Dr. Clifford Lam as accepted by Journal of Econometrics (Lam and Feng, 2018).

4.1 Introduction

In modern day finance, the so-called tick-by-tick data on the prices of financial assets are readily available together with huge volume of other financial data. Advanced computational power and efficient data storage facilities mean that these data are analyzed on a daily basis by various market makers and academic researchers. While the Markowitz portfolio theory (Markowitz, 1952) is originally proposed for a finite number of assets using inter-day price data, the easily accessible intra-day high-frequency price data for a large number assets nowadays gives rise to new possibilities for efficient portfolio allocation, on top of the apparent increase in sample size for returns and volatility matrix estimation.

Certainly, the associated challenges for using high-frequency data have to be overcome at the same time. One main challenge comes from the well documented

market microstructure noise in the recorded tick-by-tick price data (Aït-Sahalia et al., 2005; Asparouhova et al., 2013). Another challenge comes from the non-synchronous trading times when more than one asset are considered. In terms of integrated covariance estimation, Xiu (2010) suggested a maximum likelihood approach for consistent estimation under market microstructure noise. Aït-Sahalia et al. (2010) proposed a quasi-maximum likelihood approach for estimating the covariance between two assets, while Zhang (2011) proposed a two- or multi-scale covariance estimator to remove the bias accumulated due to the microstructure noise in the usual realized covariance formula, at the same time overcoming the non-synchronous trading times problem by using previous-tick times (see Chapter 4.2 also). Other attempts to overcome these two challenges together include Barndorff-Nielsen et al. (2011) and Griffin and Oomen (2011), to name but a few.

When there are more than one asset to manage, the integrated covariance matrix for the asset returns is an important input for risk management or portfolio allocation. A large number of assets requires an estimation of a large integrated covariance matrix. Even in the simplest case of independent and identically distributed random vectors, random matrix theory tells us that the sample covariance matrix will have severely biased extreme eigenvalues (see Chapter 5.2 of Bai and Silverstein (2010) for instance). To give a simple demonstration of how serious the bias problem can be, suppose we have independent and identically distributed p-dimensional random vectors $\mathbf{X} = (x_1, \ldots, x_n)^{\mathrm{T}}$ with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix. The Marčenko-Pastur Law (Marčenko and Pastur, 1967) states that the density function of the limiting spectrum of the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\text{sam}} = n^{-1}\mathbf{X}\mathbf{X}^{\mathrm{T}}$ as $p, n \to$ with $p/n \to c > 0$, is

$$p_c(x) = \begin{cases} \frac{1}{2\pi x c \sigma^2}\sqrt{(b-x)(x-a)}, & a \le x \le b; \\ 0, & \text{otherwise}, \end{cases}$$

where $a = \sigma^2(1-\sqrt{c})^2$, $b = \sigma^2(1+\sqrt{c})^2$. (See Bai and Silverstein (2010) Chapter 3.1 also.) With this, say $p = 25$ and $n = 500$, i.e., p is just 5% of n, the largest and smallest eigenvalues are 50% larger and 40% smaller than the corresponding population ones, i.e. σ^2, respectively. It means that a seemingly small p is enough already for the sample covariance matrix to suffer from significant distortion for the extreme eigenvalues, creating instability. When $\mathbf{\Sigma} \neq \sigma^2 \mathbf{I}_p$, the distortion can potentially be more severe.

To ameliorate the bias issue above, researchers propose different methods to reduce the dimension of the estimation problem, which is of order p^2, where p is the number of assets. Wang and Zou (2010), Tao et al. (2013) and Kim et al. (2016) assume a sparsity condition (perhaps after removing a market factor) and use thresholding to regularize different integrated covariance matrix estimators based on previous-tick times. Tao et al. (2011) uses a thresholded estimator to find a factor model structure for the daily dynamics of the integrated covariance matrix. These methods reduce the effective number of parameters to estimate to the order of p or less (or $p \log p$ for approximate sparsity, see Tao et al. (2013)). While consistent results are established in these methods, sparsity or factor model structure imposed regularities in the integrated covariance matrix which may not be completely satisfied in practice.

At the same time, with respect to portfolio allocation, DeMiguel et al. (2009) constrains the portfolio norm of a portfolio $\mathbf{w} = (w_1, \ldots, w_p)^{\mathrm{T}}$ using either the L_1 or squared L_2 norm, defined respectively by $\|\mathbf{w}\|_1 = \sum_i |w_i|$ and $\|\mathbf{w}\|_2^2 = \sum_i w_i^2$. Fan et al. (2012) proposes to regularize the portfolio weights by constraining the L_1 norm of the portfolio, termed the gross exposure of a portfolio in the paper. These two portfolio allocation methods do not regularize the integrated covariance matrix, but directly regularize the portfolio weights. The two-scale covariance matrix constructed in Fan et al. (2012) using the pairwise refresh method, however, may not be positive definite and adjustments are necessary to make it so. In a very broad sense, these two methods are variations of sparsity or factor model-assumed papers mentioned in the previous paragraph, essentially reducing an order p^2 problem to order p or less by assuming a sparse optimal portfolio weight.

In this chapter, we address the estimation of the integrated covariance matrix by reducing it to exactly an order p problem without assuming inherent structures to the population integrated covariance matrix or optimal portfolio weight. While this makes it impossible to estimate the integrated covariance matrix consistently, we achieve another important objective - regularization of extreme eigenvalues of the realized covariance matrix under the setting $p/n \to c > 0$ - through introducing a class of rotation-equivariant estimators and bringing it as close to the population counterpart as possible. Indeed, it is clear in our simulations and portfolio allocation exercises in Chapter 4.5 that the two-scale covariance matrix, which is essentially a realized covariance matrix, suffers from bad performance because of the instability created by the biases in its extreme eigenvalues compared to its population counterpart.

The said regularization above is achieved by minimizing a certain Frobenius error, to be discussed in Chapter 4.2.2. Such a regularization is inspired by a data splitting method originated from Abadir et al. (2014), which is proved in Lam (2016) to nonlinearly shrink the sample covariance eigenvalues at a certain data splitting ratio. We show that the resulting integrated covariance matrix estimator is consistent with a certain positive definite matrix with regularized eigenvalues at a rate of $n^{-1/6}$ under the setting $p/n \to c > 0$, with n being the sample size. This is the same rate as the univariate two-scale realized covariance estimator by Zhang (2011). We also prove the same rate of convergence when there are pervasive factors but with $p^{3/2}/n \to c > 0$. Using its inverse in the construction of the minimum variance portfolio induces a natural upper bound on the maximum exposure of the portfolio, which decays at a rate of $p^{-1/2}$ in probability when there are no pervasive factors. The importance of this bound is that the theoretical minimum variance portfolio satisfies such a bound also. See Theorem 4.3 for more details, which include results when there are pervasive factors like a market factor in the data.

The rest of the chapter is organized as follows. Chapter 4.2 presents the notations and model for the high-frequency data and introduce our way to perform nonlinear shrinkage on the two-scale covariance matrix estimator. Asymptotic theories and detailed assumptions, including those involving jumps removed data in the case of jump-diffusion log-price processes, can be found in Chapter 4.3. Practical concerns and implementation can be found in Chapter 4.4, while all simulations and a thorough empirical study are presented in Chapter 4.5. We give the conclusion in Chapter 4.6, before all the proofs of the theorems in Chapter 4.7.

4.2 Framework and Methodology

Let $\left(\Omega, \mathcal{F}, \left\{\mathcal{F}_t\right\}_{0\le t\le 1}, \mathbb{P}\right)$ be a filtered probability space on which the log-price process of the p assets under study, $\{\mathbf{X}_t\}_{0\le t\le 1}$, is adapted, where $\mathbf{X}_t = (X_t^{(1)}, \ldots, X_t^{(p)})^\intercal$. We assume $\mathbf{X}_t$ follows a diffusion process

$$d\mathbf{X}_t = \boldsymbol{\mu}_t dt + \boldsymbol{\sigma}_t d\mathbf{W}_t, \quad t \in [0,1], \tag{4.1}$$

so that the time period is normalized to have length 1. Let L be the number of partitions of the data, with

$$0 = \tau_0 < \tau_1 < \cdots < \tau_L = 1,$$

and $(\tau_{\ell-1}, \tau_\ell]$ represents the ℓth partition. The reason we partition the data is that our method of regularization is carried out within a partition at a time, with data from outside of the partition help regularize the estimator within. The partition lengths can be different, and in our case here, we set the partition as one natual day or a quarter of natual day. The ultimate estimator is then the sum of all regularized estimators for the partitions (see Chapter 4.2.2 for full details).

We assume that L is finite throughout the chapter. The process $\{\mathbf{W}_t\}$ is a p-dimensional standard Brownian motion. The drift $\boldsymbol{\mu}_t \in \mathbb{R}^p$ is random which can be correlated with $\{\mathbf{W}_t\}$. The volatility $\boldsymbol{\sigma}_t \in \mathbb{R}^{p\times p}$ is assumed to be càdlàg and non-random. For each time interval $[a, b] \subset [0, 1]$, the corresponding integrated covariance matrix is defined as

$$\boldsymbol{\Sigma}(a, b) = \int_a^b \boldsymbol{\sigma}_u \boldsymbol{\sigma}_u^{\mathrm{T}} du.$$

This matrix is an important input in risk assessment and in Markowitz portfolio allocation. If we have a portfolio $\mathbf{w}$ which stays constant over a period of time $[a, b]$, then the risk of the portfolio over this period of time can be expressed as

$$R^{1/2}(\mathbf{w}) = (\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}(a, b)\mathbf{w})^{1/2} = \Big(\int_a^b \mathbf{w}^{\mathrm{T}}\boldsymbol{\sigma}_t\boldsymbol{\sigma}_t^{\mathrm{T}}\mathbf{w}\, dt\Big)^{1/2}.$$

The integrand $\mathbf{w}^{\mathrm{T}}\boldsymbol{\sigma}_t\boldsymbol{\sigma}_t^{\mathrm{T}}\mathbf{w}$ can be considered an instantaneous squared risk at time t for $\mathbf{w}$, and hence $R(\mathbf{w})$ is a measure of the total risk accumulated over the period $[a, b]$. At the same time, in Markowitz portfolio allocation for instance, $\boldsymbol{\Sigma}(a, b)^{-1}$ is required for the construction of the minimum variance portfolio (see Chapter 4.3.2 for more details).

Let $\{v_s\}, 1 \le s \le nL$ be the set of all-refresh times for the log prices in $\mathbf{X}_t$, where $n(\ell)$ is the number of all-refresh times at partition ℓ, $\ell = 1, \ldots, L$, and $n = L^{-1}\sum_{\ell=1}^{L} n(\ell)$ is the average number of all-refresh times in a partition, which has the same order as the total sample size nL since L is finite. To recall, an all-refresh time v_s is the time when all assets have been traded at least once from the last all-refresh time v_{s-1}. Let $t_s^j \in (v_{s-1}, v_s]$ be the sth previous-tick time for the jth asset, which is the last trading time before or at v_s. For non-synchronous trading, $t_s^{j_1} \neq t_s^{j_2}$ for $j_1 \neq j_2$ in general. Also,

high-frequency prices are typically contaminated by microstructure noise, so that at the all-refresh time v_s, we only observe

$$\mathbf{Y}(s) = \mathbf{X}(s) + \boldsymbol{\epsilon}(s), \ s = 1, \ldots, nL, \tag{4.2}$$

where $\mathbf{X}(s) = (X^{(1)}_{t^1_s}, \ldots, X^{(p)}_{t^p_s})^\top$ and $\boldsymbol{\epsilon}(s) = (\epsilon^{(1)}_{t^1_s}, \ldots, \epsilon^{(p)}_{t^p_s})^\top$, and $\boldsymbol{\epsilon}(\cdot)$ can be dependent on $\mathbf{X}(\cdot)$ in general (see the assumptions in Chapter 4.3). The underlying microstructure noise process $\{\boldsymbol{\epsilon}_t\}_{0\le t\le 1}$ is assumed to be adapted to $\{\mathcal{F}_t\}_{0\le t\le 1}$, so that the observed price process $\{\mathbf{Y}_t\}_{0\le t\le 1}$ is also adapted.

4.2.1 Two-Scale Covariance Estimator

Contamination of microstructure noise in high-frequency data means that the usual realized covariance is heavily biased. Hence in Zhang (2011), a Two-Scale CoVariance estimator (TSCV) is introduced to remove this bias. In this chapter, we use a slightly modified multivariate version of the two-scale covariance estimator, also by Zhang (2011). For $\ell = 1, \ldots, L$, define

$$\begin{aligned}
\widehat{\langle \mathbf{Y}, \mathbf{Y}^\top \rangle}_\ell &= [\mathbf{Y}, \mathbf{Y}^\top]^{(K)}_\ell - \frac{|S^\ell(K)|_K}{|S^\ell(1)|}[\mathbf{Y}, \mathbf{Y}^\top]^{(1)}_\ell, \text{ with} \\
\left([\mathbf{Y}, \mathbf{Y}^\top]^{(m)}_\ell\right)_{i,j} &= [Y^{(i)}, Y^{(j)}]^{(m)}_\ell = \frac{1}{m}\sum_{r\in S^\ell(m)} (Y^{(i)}_{t^i_r} - Y^{(i)}_{t^i_{r-m}})(Y^{(j)}_{t^j_r} - Y^{(j)}_{t^j_{r-m}}), \text{ and} \\
S^\ell(m) &= \{r : t^i_r, t^i_{r-m} \in (\tau_{\ell-1}, \tau_\ell] \text{ for all } i\}, \ |S^\ell(m)|_m = \frac{|S^\ell(m)| - m + 1}{m}.
\end{aligned} \tag{4.3}$$

Here $|S(m)|$ denotes the number of elements in S while $|S(m)|_m$ represents the adjusted number of elements. Note that $[Y^{(i)}, Y^{(j)}]^{(1)}_\ell$ is the usual realized covariance matrix when returns are calculated using adjacent previous-tick times, whereas $[Y^{(i)}, Y^{(j)}]^{(K)}_\ell$ can be seen as a realized covariance matrix when returns are calculated at time points which are K previous-tick times apart instead of 1, i.e. another scale. Ultimately, while both are dominated by the market microstructure noise, the difference defined in $\widehat{\langle \mathbf{Y}, \mathbf{Y}^\top \rangle}_\ell$ is proved in Zhang (2011) to be able to cancel out the dominating effect of the microstructure noise. With this, we define TSCV for the partition $(\tau_{\ell-1}, \tau_\ell]$ to be

$$\widetilde{\boldsymbol{\Sigma}}(\tau_{\ell-1}, \tau_\ell) = \widehat{\langle \mathbf{Y}, \mathbf{Y}^\top \rangle}_\ell. \tag{4.4}$$

We suppress the dependence on K in the notation $\widetilde{\mathbf{\Sigma}}(\tau_{\ell-1}, \tau_\ell)$ and all related definitions in the next chapter. In Chapter 4.3, we show that K works well at the order $n^{2/3}$, which is indeed the order of magnitude suggested in Zhang (2011).

Remark 4.1 *The Multi-Scale Realized Volatility Matrix (MSRVM) by Tao et al. (2013), the Kernel Realized Volatility Matrix (KRVM) by Barndorff-Nielsen et al. (2011) and the Pre-averaging Realized Volatility Matrix (PRVM) by Christensen et al. (2010) all have better convergence rates than the TSCV for multivariate settings. The latter two estimators can be constructed to be positive semi-definite, although all three estimators do not allow p to be growing with n. In principle, our regularized estimator, to be introduced in Chapter 4.2.2, can be based on regularizing these three estimators. However, while the proof of our regularization method on the MSRVM is an extension of ours on the TSCV (because MSRVM involves sums of order of $n^{1/2}$ terms), the jittering and pre-averaging operations on the KRVM and PRVM respectively are more difficult to handle in the proofs. We decide to leave the extensions of our regularization method to these estimators in a future project.*

4.2.2 Our Proposed Integrated Covariance Matrix Estimator

Although the two-scale covariance estimator in equation (4.4) removes the bias contributed from the microstructure noise, it does not solve the bias issue for the extreme eigenvalues when p is large such that $p/n \to c > 0$, where the spread of the eigenvalues in the realized covariance matrix $\widetilde{\mathbf{\Sigma}}(\tau_{j-1}, \tau_j)$ is much larger than the population counterpart, creating instability. In a setting with stationary covariance matrix, Abadir et al. (2014) introduced the idea of splitting the data into two parts in order to regularize the sample covariance matrix constructed from one part of the data. Lam (2016) showed that with a certain splitting ratio, in fact the extreme eigenvalues of the sample covariance matrix are nonlinear shrunk asymptotically, the same as the nonlinear shrinkage introduced in Ledoit and Wolf (2012). We employ the data splitting idea in Abadir et al. (2014) for our high-frequency data setting in this chapter. In order to regularize the realized covariance matrix in the time period $(\tau_{j-1}, \tau_j]$, $j = 1, \ldots, L$, we follow Lam (2016) and consider a rotation-equivariant estimator $\mathbf{\Sigma}(\mathbf{D}) = \mathbf{P}_{-j}\mathbf{D}\mathbf{P}_{-j}^{\mathrm{T}}$, where $\mathbf{D}$ is a diagonal matrix, and $\mathbf{P}_{-j}$ is orthogonal such that

$$\widetilde{\mathbf{\Sigma}}_{-j} = \mathbf{P}_{-j}\mathbf{D}_{-j}\mathbf{P}_{-j}^{\mathrm{T}}, \quad j = 1, \ldots, L, \quad \text{with } \widetilde{\mathbf{\Sigma}}_{-j} = \sum_{\ell \neq j} \widetilde{\mathbf{\Sigma}}(\tau_{\ell-1}, \tau_\ell). \tag{4.5}$$

The class of rotation-equivariant estimators allows for the same rotation of the estimator when the observed vectors are rotated. This is first introduced in James and Stein (1961) for estimating a covariance matrix under the Stein's loss function, with respect to which this class is invariant under rotation. Hence with no *a priori* information of the eigenvectors of the population covariance matrix, this class provides a good starting point as an estimator. Ledoit and Wolf (2012) used this class of estimators for the purpose of nonlinear shrinkage of eigenvalues. However, high-frequency data vectors are in general not independent and identically distributed, so that the explicit nonlinear shrinkage formula in Ledoit and Wolf (2012) cannot be used.

To introduce our estimator, consider the following optimization problem, with similar problem considered in Ledoit and Wolf (2012) and Lam (2016):

$$\min_{\mathbf{D}\ \text{diagonal}} \|\mathbf{P}_{-j}\mathbf{D}\mathbf{P}_{-j}^{\mathrm{T}} - \boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\|_F, \tag{4.6}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Unlike Ledoit and Wolf (2012) which uses the eigenmatrix of the sample covariance constructed from the full data set, we use $\mathbf{P}_{-j}$ for the rotation-equivariant class. This facilitates regularization by allowing us to condition on the information outside of partition j, which weaken the correlation between $\{\mathbf{X}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$, and the serial correlation within $\{\boldsymbol{\epsilon}_t\}$ (see Assumption (E3) in Chapter 4.3).

Proposition 4.1 *The optimization problem (4.6) has solution* $\mathbf{D} = \mathrm{diag}(\mathbf{P}_{-j}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{P}_{-j})$, *where* $\mathrm{diag}(A)$ *creates a diagonal matrix using the diagonal elements of* A.

Proof of Proposition 4.1. To simplify notations in this proof, write $\mathbf{D} = \mathrm{diag}(d_1,\ldots,d_p)$, $\mathbf{P}_{-j} = \mathbf{Q} = (\mathbf{q}_1,\ldots,\mathbf{q}_p)$ and $\boldsymbol{\Sigma}(\tau_{j-1},\tau_j) = \boldsymbol{\Sigma}_j$. Then

$$\begin{aligned}\|\mathbf{P}_{-j}\mathbf{D}\mathbf{P}_{-j}^{\mathrm{T}} - \boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\|_F^2 &= \mathrm{tr}(\mathbf{D} - \mathbf{Q}^{\mathrm{T}}\boldsymbol{\Sigma}_j\mathbf{Q})^2 = \sum_{i=1}^{p} d_i^2 - 2\mathrm{tr}(\mathbf{D}\mathbf{Q}^{\mathrm{T}}\boldsymbol{\Sigma}_j\mathbf{Q}) + \mathrm{tr}(\mathbf{Q}^{\mathrm{T}}\boldsymbol{\Sigma}_j^2\mathbf{Q}) \\ &= \sum_{i=1}^{p} d_i^2 - 2\sum_{i=1}^{p} d_i\mathbf{q}_i^{\mathrm{T}}\boldsymbol{\Sigma}_j\mathbf{q}_i + \mathrm{tr}(\mathbf{Q}^{\mathrm{T}}\boldsymbol{\Sigma}_j^2\mathbf{Q}).\end{aligned}$$

Differentiating the above with respect to d_i and set the derivative to 0, we get $d_i = \mathbf{q}_i^{\mathrm{T}}\boldsymbol{\Sigma}_j\mathbf{q}_i$, which leads to the solution $\mathbf{D} = \mathrm{diag}(\mathbf{P}_{-j}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{P}_{-j})$. □

Clearly, all eigenvalues of $\mathbf{D}$ are contained within the largest and smallest eigenvalues of $\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)$. This way, the spread of the eigenvalues in $\mathbf{D}$ is regularized. Ultimately,

we can prove that all the elements in $\text{diag}(\mathbf{P}_{-j}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}(\tau_{j-1},\tau_j)\mathbf{P}_{-j})$ are asymptotically close to those in $\mathbf{D} = \text{diag}(\mathbf{P}_{-j}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{P}_{-j})$ in probability (see Theorem 4.1). This allows us to define our integrated covariance matrix estimator for the partition $(\tau_{j-1},\tau_j]$ to be

$$\widehat{\mathbf{\Sigma}}(\tau_{j-1},\tau_j) = \mathbf{P}_{-j}\text{diag}(\mathbf{P}_{-j}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}(\tau_{j-1},\tau_j)\mathbf{P}_{-j})\mathbf{P}_{-j}^{\mathrm{T}}. \tag{4.7}$$

The overall integrated covariance matrix estimator for the period $[0,1]$ is then defined to be

$$\widehat{\mathbf{\Sigma}}(0,1) = \sum_{j=1}^{L}\widehat{\mathbf{\Sigma}}(\tau_{j-1},\tau_j) = \sum_{j=1}^{L}\mathbf{P}_{-j}\text{diag}(\mathbf{P}_{-j}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}(\tau_{j-1},\tau_j)\mathbf{P}_{-j})\mathbf{P}_{-j}^{\mathrm{T}}. \tag{4.8}$$

In using $\mathbf{P}_{-j}$, we assume that each interval $(\tau_{j-1},\tau_j]$ is small, so that at the population level, the eigenvectors of each $\mathbf{\Sigma}_{-j} = \sum_{\ell\neq j}\mathbf{\Sigma}(\tau_{\ell-1},\tau_\ell)$ is not far from those for $\mathbf{\Sigma}(0,1)$. Then each $\mathbf{P}_{-j}$ should also be similar to the eigenmatrix $\mathbf{P}$ for $\widetilde{\mathbf{\Sigma}}(0,1) = \sum_{\ell=1}^{L}\widetilde{\mathbf{\Sigma}}(\tau_{\ell-1},\tau_\ell)$. The estimator $\widehat{\mathbf{\Sigma}}(0,1)$ in (4.8) is then

$$\begin{aligned}\widehat{\mathbf{\Sigma}}(0,1) &\approx \sum_{j=1}^{L}\mathbf{P}_{-j}\text{diag}(\mathbf{P}_{-j}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{P}_{-j})\mathbf{P}_{-j}^{\mathrm{T}}\\ &\approx \mathbf{P}\text{diag}\left(\mathbf{P}^{\mathrm{T}}\sum_{j=1}^{L}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{P}\right)\mathbf{P}^{\mathrm{T}} = \mathbf{P}\text{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{\Sigma}(0,1)\mathbf{P})\mathbf{P}^{\mathrm{T}}.\end{aligned}$$

The first approximation uses the result in Theorem 4.1 to be presented in Chapter 4.3 below. The estimator $\mathbf{P}\text{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{\Sigma}(0,1)\mathbf{P})\mathbf{P}^{\mathrm{T}}$ can be considered as an ideal rotation-equivariant estimator, where $\mathbf{P}$ is an eigenmatrix utilizing all of the all-refresh data points, and $\text{diag}(\mathbf{P}^{\mathrm{T}}\mathbf{\Sigma}(0,1)\mathbf{P})$ is the ideal diagonal matrix in terms of the Frobenius error.

In practice, a small interval can be a trading day or a quarter of it, depending on the number of trading days of data we have and the number of all-refresh data points in them. We propose an optimization criterion to choose the number of partitions (not necessarily uniform) in Chapter 4.4. In our simulations and empirical examples in Chapter 4.5, we use 5 or 1 day of training data with $(\tau_{\ell-1},\tau_\ell]$ set as 1 day or a quarter of a day, with the number of all-refresh data points in the order of hundreds in each interval.

4.3 Asymptotic Theory

In this chapter, we show that our proposed estimator (4.7) in the jth partition of the data is asymptotically close to the corresponding ideal rotation-equivariant estimator

$$\mathbf{\Sigma}_{\text{Ideal}}(\tau_{j-1},\tau_j) = \mathbf{P}_{-j}\text{diag}(\mathbf{P}_{-j}^{\text{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{P}_{-j})\mathbf{P}_{-j}^{\text{T}}. \tag{4.9}$$

This is exactly the optimal estimator that solves (4.6) following Proposition 4.1. While $\widetilde{\mathbf{\Sigma}}(\tau_{j-1},\tau_j)$ can have its spread of eigenvalues much larger than that of $\mathbf{\Sigma}(\tau_{j-1},\tau_j)$ when $p/n \to c > 0$, our estimator $\widehat{\mathbf{\Sigma}}(\tau_{j-1},\tau_j)$ in equation (4.7) has its spread of eigenvalues contained within the spread of $\mathbf{\Sigma}(\tau_{j-1},\tau_j)$ asymptotically by being close to $\mathbf{\Sigma}_{\text{Ideal}}(\tau_{j-1},\tau_j)$ in equation (4.9) above (see Theorem 4.1 below). We first introduce some assumptions for our theorems to hold. We write $a \asymp b$ to mean that $a = O(b)$ and $b = O(a)$, and $a \asymp_P b$ to mean that $a = O_P(b)$ and $b = O_P(a)$.

For $j = 1,\ldots,L$, and $v_s = v_s^j$ which is the sth all-refresh time within partition j, define

$$\mathcal{F}_{-j} = \mathcal{F}_{\tau_{j-1}} \cup \mathcal{F}/\mathcal{F}_{\tau_j}, \quad \mathcal{F}_s^j = \mathcal{F}_{v_s}/\mathcal{F}_{\tau_{j-1}},$$

with $\mathcal{F}_s^j = \emptyset$ for $s \le 0$. The following assumptions are true for $K = 1$ or $K \asymp n^{2/3}$.

Assumptions on the drift $\boldsymbol{\mu}_t$:

(D1) The drift $\boldsymbol{\mu}_t$ has càdlàg components and is random, such that for $s = K, K+1,\ldots,n(j)$,

$$\int_{v_{s-K}}^{v_s} \boldsymbol{\mu}_t dt = \mathbf{A}(v_{s-K},v_s)\mathbf{Z}_{d,s}^j,$$

where $\mathbf{A}(v_{s-K},v_s) \neq \mathbf{0}$ is a non-random $p \times p$ matrix and can be asymmetric and singular. It has $\|\mathbf{A}(v_{s-K},v_s)\| = O(p^{1/2}K^{1/2}|v_s - v_{s-1}|)$, where the order $p^{1/2}$ only appears when there are only finite number of columns (say r) that are non-zero. The random vector $\mathbf{Z}_{d,s}^j \in \mathcal{F}_s^j$ has components conditionally independent of each other given $\mathcal{F}_{-j}$, with eighth moments exist. Also, $E(\mathbf{Z}_{d,s}^j|\mathcal{F}_{-j}) = \mathbf{0}$ and $\text{var}(\mathbf{Z}_{d,s}^j|\mathcal{F}_{-j}) = \mathbf{I}_p$ almost surely.
The drift term $\boldsymbol{\mu}_t$ can also be non-random, in which case $\mathbf{Z}_{d,s}^j = (1,0,\ldots,0)^{\text{T}}$ for small s, and the assumption for $\mathbf{A}(v_{s-K},v_s)$ is the same as above.

(D2) Write $\mathbf{P}_{-j} = (\mathbf{p}_{1j},\ldots,\mathbf{p}_{pj})$. We assume for each $i = 1,\ldots,p$, and $s = rK+q$ for $r = 1,\ldots,|S^j(K)|_K$ and $q = 0,1,\ldots,K-1$, there exists $\rho_{d,K,q}^j \in \mathcal{F}_{-j}$ such that

$0 \leq \rho^j_{d,K,q} \leq \xi < 1$ with ξ a constant, and for $\ell = K+q, 2K+q, \ldots, rK+q$,

$$E\Big((\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K},v_s)\mathbf{Z}^j_{d,\ell})^2|\mathcal{F}_{-j}\cup\mathcal{F}^j_{\ell-K}\Big)$$
$$= \rho^j_{d,K,q}(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K},v_s)\mathbf{Z}^j_{d,\ell-K})^2 + (1-\rho^j_{d,K,q})\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K},v_s)\mathbf{A}(v_{s-K},v_s)^{\mathrm{T}}\mathbf{p}_{ij} + e^{ij}_{d,\ell-K},$$

where we define $\mathbf{Z}^j_{d,\ell}\mathbf{Z}^{j\mathrm{T}}_{d,\ell} = \mathbf{I}_p$ and $e^{ij}_{d,\ell} = 0$ for $\ell \leq 0$. The process $\{e^{ij}_{d,\ell}\}$ with $e^{ij}_{d,\ell} \in \mathcal{F}^j_\ell$ has $E(e^{ij}_{d,\ell}|\mathcal{F}_{-j}\cup\mathcal{F}^j_{\ell-K}) = 0$ almost surely, and $e^{ij}_{d,\ell}|\mathcal{F}_{-j}\cup\mathcal{F}^j_{\ell-K} = O_P(\|\mathbf{A}(v_{s-K},v_s)\|^2)$.

(D3) Let $\psi(x) = e^{x^2} - 1$. We assume that for $\ell = 0, 1, \ldots, s$,

$$E\left\{\psi\left(\frac{|(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K},v_s)\mathbf{Z}^j_{d,\ell})^2 - \mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K},v_s)\mathbf{A}(v_{s-K},v_s)^{\mathrm{T}}\mathbf{p}_{ij}|}{(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K},v_s)\mathbf{Z}^j_{d,\ell-K})^2 + \|\mathbf{A}(v_{s-K},v_s)\|^2}\right)\Big|\mathcal{F}_{-j}\cup\mathcal{F}^j_{\ell-K}\right\} < \infty,$$
$$E\left\{\psi\left(\frac{|e^{ij}_{d,\ell}|}{(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K},v_s)\mathbf{Z}^j_{d,\ell-K})^2 + \|\mathbf{A}(v_{s-K},v_s)\|^2}\right)\Big|\mathcal{F}_{-j}\cup\mathcal{F}^j_{\ell-K}\right\} < \infty.$$

Assumptions on the volatility $\boldsymbol{\sigma}_t$ and Brownian motion $\mathbf{W}_t$:

(V1) The volatility $\boldsymbol{\sigma}_t$ has càdlàg components and is non-random, and the Brownian motion $\{\mathbf{W}_t\}$ can be correlated with $\{\boldsymbol{\mu}_t\}$ in general. Write

$$\int_{v_{s-K}}^{v_s} \boldsymbol{\sigma}_t d\mathbf{W}_t = \boldsymbol{\Sigma}(v_{s-K},v_s)^{1/2}\mathbf{Z}^j_{v,s},$$

where $\boldsymbol{\Sigma}(v_{s-K},v_s)$ is a symmetric positive definite $p\times p$ matrix which can be random, with

$$\lambda_{\min}(\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)) \geq C(\tau_j-\tau_{j-1})^{-1},\ \lambda_{\max}(\boldsymbol{\Sigma}(v_{s-K},v_s)) \asymp_P \|\mathbf{A}(v_{s-K},v_s)\|^2/|v_s-v_{s-K}|,$$

where $C > 0$ is a constant. The process $\{\boldsymbol{\sigma}_t\}$ is independent of all other processes. Also, $E(\mathbf{Z}^j_{v,s}|\mathcal{F}_{-j}) = \mathbf{0}$ and $\mathrm{var}(\mathbf{Z}^j_{v,s}|\mathcal{F}_{-j}) = \mathbf{I}_p$ almost surely. The random vector $\mathbf{Z}^j_{v,s} \in \mathcal{F}^j_s$ has components conditionally independent of each other given $\mathcal{F}_{-j}$, with eighth moments exist.

(V2) Parallel to (D2), but expectations are taken conditional on $\mathcal{F}_{-j}\cup\mathcal{F}^j_{\ell-K}\cup\mathcal{F}^{\boldsymbol{\sigma}}_{v_s}$, where $\mathcal{F}^\sigma_t$ is the σ-algebra generated by the process $\{\boldsymbol{\sigma}_t\}$ up to time t.
Also, $\rho^j_{d,K,q}$ replaced by $\rho^j_{v,K,q} \in \mathcal{F}_{-j}$, $\mathbf{A}(v_{s-K},v_s)$ by $\boldsymbol{\Sigma}(v_{s-K},v_s)^{1/2}$, $\mathbf{Z}^j_{d,\ell}$ by $\mathbf{Z}^j_{v,\ell}$ and $e^{ij}_{d,\ell}$ by $e^{ij}_{v,\ell}$ with $e^{ij}_{v,\ell}|\mathcal{F}_{-j}\cup\mathcal{F}^j_{\ell-K}\cup\mathcal{F}^{\boldsymbol{\sigma}}_{v_s} = O_P(e^{ij}_{d,\ell}|\mathcal{F}_{-j}\cup\mathcal{F}^j_{\ell-K})/|v_s - v_{s-K}|$.

(V3) Parallel to (D3), replacements the same as in (V2).

Assumptions on the microstructure noise $\boldsymbol{\epsilon}_t$:

(E1) Within the jth partition, $E(\boldsymbol{\epsilon}(s)\boldsymbol{\epsilon}(s)^{\mathrm{T}}|\mathcal{F}_{-j}) = \boldsymbol{\Sigma}^j_{\epsilon,s}$, which is random and independent of all other processes given $\mathcal{F}_{-j}$. Also, $E(\boldsymbol{\Sigma}^j_{\epsilon,s}) = \boldsymbol{\Sigma}^j_{\epsilon}$, and $\|\boldsymbol{\Sigma}^j_{\epsilon,s}\| \leq \lambda_\epsilon < \infty$ uniformly as $n, p \to \infty$ where λ_ϵ is a constant.

(E2) Within the jth partition, we can write $\boldsymbol{\epsilon}(s) = (\boldsymbol{\Sigma}^j_{\epsilon,s})^{1/2}\mathbf{Z}^j_{\epsilon,s}$, with $\mathbf{Z}^j_{\epsilon,s} \in \mathcal{F}^j_s$ having conditionally independent components given $\mathcal{F}_{-j}$. Also $E(\mathbf{Z}^j_{\epsilon,s}|\mathcal{F}_{-j}) = \mathbf{0}$ almost surely and eighth order moments exist for the components of $\mathbf{Z}^j_{\epsilon,s}$.

(E3) Let $\mathcal{F}^X_t$ be the σ-algebra generated by the log-price processes up to time t, and $\mathcal{F}^\epsilon_t$ the one by the microstructure noise processes up to time t, so that $\mathcal{F}_t = \bigcap_{s>t} \mathcal{F}^X_s \otimes \mathcal{F}^\epsilon_s$. Then for s_1, s_2 time points within partition j, given $\mathcal{F}_{-j}$, we assume the φ-mixing coefficient between two σ-algebras satisfies

$$\varphi(\mathcal{F}^X_{s_1}, \mathcal{F}^\epsilon_{s_2}|\mathcal{F}_{-j}) = O(n^{-1}) = \varphi(\mathcal{F}^\epsilon_{s_2}, \mathcal{F}^X_{s_1}|\mathcal{F}_{-j}).$$

Also, for $s_2 > s_1$ time points within partition j, we assume

$$\varphi(\mathcal{F}^\epsilon_{s_1}, \mathcal{F}^\epsilon_{s_2}/\mathcal{F}^\epsilon_{s_1}|\mathcal{F}_{-j}) = O(n^{-1}) = \varphi(\mathcal{F}^\epsilon_{s_2}/\mathcal{F}^\epsilon_{s_1}, \mathcal{F}^\epsilon_{s_1}|\mathcal{F}_{-j}).$$

Other assumptions:

(O1) The observation times are independent of $\mathbf{X}(\cdot)$ and $\boldsymbol{\epsilon}(\cdot)$, and the partition boundaries τ_ℓ, $\ell = 0, 1, \ldots, L$, satisfy $0 < C_1 \leq \min_{\ell=1,\ldots,L} L(\tau_\ell - \tau_{\ell-1}) \leq \max_{\ell=1,\ldots,L} L(\tau_\ell - \tau_{\ell-1}) \leq C_2 < \infty$, where C_1, C_2 are generic constants. Also, the all-refresh times v_s, $s = 1, \ldots, nL$ satisfy $\max_{s=1,\ldots,nL} nL(v_s - v_{s-1}) \leq C_3$ for a generic constant $C_3 > 0$. Moreover, $\max_{\ell=1,\ldots,L} nL(\tau_\ell - v_{n(\ell)}) = o(1)$. The sample size in the jth partition has $n(j)/n \to 1$.

O2) The pervasive factors, if any, persist within an interval $(v_{s-1}, v_s]$ for $s = 1, \ldots, nL$.

There is another set of assumptions (O3) to (O5) in Chapter 4.7. They involve the drift and volatility in $\mathbf{X}_{v_s} - \mathbf{X}(s)$, i.e. the drift and volatility in between the all-refresh and the previous-tick times. These assumptions are in many ways parallel to assumptions

(D1) to (D3) and (V1) to (V3), but the decompositions are more involved, so that we choose to present them in Chapter 4.7 to aid the flow of the chapter.

The matrix $\mathbf{A}(v_{s-K}, v_s)$ in assumptions (D1) to (D3) plays the role of a factor loading matrix in a factor model if the drift $\boldsymbol{\mu}_t$ is random. Within partition j, if $\mathbf{A}(v_{s-K}, v_s)$ is diagonal, the contribution of drift among all assets over v_{s-K} to v_s are conditionally independent given $\mathcal{F}_{-j}$. If $\mathbf{A}(v_{s-K}, v_s)$ is singular with only the first $r \ll p$ columns being non-zero, then it represents an exact r-factor model with no noise on the drift. The first r singular values of $\mathbf{A}(v_{s-K}, v_s)$ are then of order $p^{1/2}K^{1/2}|v_s - v_{s-1}|$, with $K^{1/2}|v_s - v_{s-1}|$ accounting for the length of the time interval considered.

The serial dependence of the drift vector is depicted in Assumption (D2). This assumption is more general than it seems. For instance, $\mathbf{Z}^j_{d,\ell}$ can be a random vector of maringales, so that

$$E(\mathbf{Z}^j_{d,\ell}|\mathcal{F}_{-j} \cup \mathcal{F}^j_{\ell-K}) = \mathbf{Z}^j_{d,\ell-K},$$

and hence $E(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K}, v_s)\mathbf{Z}^j_{d,\ell}|\mathcal{F}_{-j} \cup \mathcal{F}^j_{\ell-K}) = \mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K}, v_s)\mathbf{Z}^j_{d,\ell-K}$. Then by Jensen's inequality,

$$E((\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K}, v_s)\mathbf{Z}^j_{d,\ell})^2|\mathcal{F}_{-j} \cup \mathcal{F}^j_{\ell-K}) \geq (\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K}, v_s)\mathbf{Z}^j_{d,\ell-K})^2,$$

and the assumption only requires a uniformly strict inequality above, so that $\rho^j_{d,K,q}$ can be uniformly smaller than 1. Note also

$$E((\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K,v_s})\mathbf{Z}^j_{d,\ell})^2|\mathcal{F}_{-j}) = \mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K}, v_s)\mathbf{A}(v_{s-K}, v_s)^{\mathrm{T}}\mathbf{p}_{ij},$$

and hence the assumption blances this mean with the squared conditional expected value of the martingale, subject to an error $e^{ij}_{d,\ell}$.

If $\mathbf{Z}^j_{d,\ell}$ is independent of any past information such that $E(\mathbf{Z}^j_{d,\ell}|\mathcal{F}_{-j} \cup \mathcal{F}^j_{\ell-K}) = \mathbf{0}$, then

$$E((\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K,v_s})\mathbf{Z}^j_{d,\ell})^2|\mathcal{F}_{-j} \cup \mathcal{F}^j_{\ell-K}) = \mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K}, v_s)\mathbf{A}(v_{s-K}, v_s)^{\mathrm{T}}\mathbf{p}_{ij},$$

so that Assumption (D2) means that $\rho^j_{d,K,q} = e^{ij}_{d,\ell} = 0$.

Assumption (D3) says that quadratic forms not too far in time apart can be very different but with sub-Gaussian-tailed probability. Assumptions (D1) to (D3) together allow us to use certain Hoeffding's inequalities for sums of martingale differences (see van de Geer (2002), Theorem 2.2).

If the drift $\boldsymbol{\mu}_t$ is non-random, then the matrix $\mathbf{A}(v_{s-K}, v_s)$ can be set as zero except the first column which is a non-zero known vector. With $\mathbf{Z}_{d,s}^j = (1, 0, \ldots, 0)^{\mathrm{T}}$, assumptions (D2) and (D3) are automatically satisfied with $e_{d,\ell}^{ij} = 0$. We do not make further assumptions for $\mathbf{A}(\cdot, \cdot)$, and hence the drift can include longer term trends (where components of $\mathbf{A}(\cdot, \cdot)$ can be increasing or decreasing over different tie segments) and pervasive factors.

Assumption (V1) to (V3) for the volatility are parallel to (D1) to (D3). The subtler part is in Assumption (V1), where $\|\boldsymbol{\Sigma}(v_{s-K}, v_s)\|$ depends on $\|\mathbf{A}(v_{s-K}, v_s)\|$. In doing so, we are essentially assuming that if there are pervasive factors such as the market factor, then they affect both the drift and the volatility of the log-price process at the same time, which certainly makes sense. Then order $p^{1/2}$ singular values in $\mathbf{A}(v_{s-K}, v_s)$ translates to order p eigenvalues in $\boldsymbol{\Sigma}(v_{s-K}, v_s)$ in the presence of pervasive factors, appropriately adjusted by $|v_s - v_{s-K}|$.

Assumption (E1) allows for time-varying covariance matrix for the microstructure noise. Assumption (E3) particularly assumes a weak dependence between the log-price process and the microstructure noise process within partition j, as well as a weak serial dependence among the microstructure noise vectors, when $\mathcal{F}_{-j}$ is given. This assumption is inspired by Chen and Mykland (2017), where they assumed that given the entire information of the log-price process, the microstructure noise at different time points are independent. In our case, we are not given the entire picture of the log-price process, but not far from that either since with $\mathcal{F}_{-j}$ we are given $nL - n(j)$ data points from the total of nL. Then instead of assuming the microstructure noise vectors are independent, we assume that they are weakly dependent, and with n larger (i.e., more information at more time points available outside partition j), the dependence is weaker.

The first part of Assumption (O1) is automatically satisfied if the boundary set $\{\tau_\ell\}_{0 \le \ell \le L}$ is pre-set, for instance, to be the daily opening or closing time of the L days of data, or a quarter of it, just as described in Chapter 4.2.2. See also Chapter 4.4 on a criterion in choosing these tuning parameters. Assumption (O2) means that the pervasive factors are either present between two all-refresh times, or they are absent.

Theorem 4.1 *Let Assumptions (D1) to (D3), (V1) to (V3), (E1) to (E3) and (O1) to (O5) hold. For the all-refresh log-price data* $\mathbf{Y}(s)$, $s = 1, \ldots, nL$ *in (4.2), as* $n, p \to \infty$ *such that* $p/n \to c > 0$, *if there are no pervasive factors, i.e.* $\|\mathbf{A}(v_{s-K}, v_s)\| =$

$O(K^{1/2}|v_s - v_{s-1}|)$, the integrated covariance matrix estimator constructed in (4.7) and $\widehat{\mathbf{\Sigma}}(0,1)$ in (4.8) satisfy

$$\max_{j=1,\ldots,L} \|\widehat{\mathbf{\Sigma}}(\tau_{j-1},\tau_j)\mathbf{\Sigma}_{\text{Ideal}}(\tau_{j-1},\tau_j)^{-1} - \mathbf{I}_p\| = O_P(n^{-1/6}),$$
$$\|\widehat{\mathbf{\Sigma}}(0,1)\mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1} - \mathbf{I}_p\| = O_P(n^{-1/6}),$$

where $\|\cdot\|$ denotes the spectral norm of a matrix. If there are pervasive factors so that $\|\mathbf{A}(v_{s-K},v_s)\| = O(p^{1/2}K^{1/2}|v_s - v_{s-1}|)$ (this includes the case when $\boldsymbol{\mu}_t$ is assumed non-random), then assuming $p^{3/2}/n \to c > 0$, the above results still hold.

The proof can be found in Chapter 4.7. The rate of convergence of our estimator is $n^{-1/6}$, the same as the TSCV in the univariate case (Zhang, 2011). Note that Assumptions (D1) and (V1) allow for the existence of pervasive factors like the market factor, and our estimator is still converging to the ideal estimator in probability at a rate of $n^{-1/6}$ if $p^{3/2}/n \to c > 0$. One remarkable fact is that this rate does not depend on p. We require p to be growing slower than n in the presence of pervasive factors mainly because the drift term that can overwhelm the estimator when there are pervasive factors. When the drift is non-random under Assumption (D1), it certainly can behave as if there are pervasive factors when there are no further assumptions on $\mathbf{A}(\cdot,\cdot)$, and we do need $p^{3/2}/n \to c > 0$ for the results in Theorem 4.1 to hold (see Remark 4.2 as well). Indeed, without a drift term, Lam (2016) allows the (low frequency) data to have a factor structure under $p/n \to c > 0$.

Since $\mathbf{P}_{-j}$ is orthogonal, it is easy to see that $\mathbf{\Sigma}_{\text{Ideal}}(\tau_{j-1},\tau_j)$ in (4.9) has

$$\text{Cond}(\mathbf{\Sigma}_{\text{Ideal}}(\tau_{j-1},\tau_{j-1})) \le \text{Cond}(\mathbf{\Sigma}(\tau_{j-1},\tau_j)),$$

where $\text{Cond}(\cdot)$ is the condition number of a matrix, defined by dividing the maximum over the minimum magnitude of eigenvalue of the matrix. Theorem 4.1 then implies that

$$\text{Cond}(\widehat{\mathbf{\Sigma}}(\tau_{j-1},\tau_{j-1})) \le \text{Cond}(\mathbf{\Sigma}(\tau_{j-1},\tau_j))$$

in probability. This is the result of nonlinear shrinkage of the eigenvalues in $\widehat{\mathbf{\Sigma}}(\tau_{j-1},\tau_j)$. Our estimator then has its spread of eigenvalues contained within the population counterpart, so that it is more stable than $\widetilde{\mathbf{\Sigma}}(\tau_{j-1},\tau_j)$, which can have its extreme eigenvalues sverely biased when $p/n \to c > 0$, creating instability. The TSCV indeed

performs worse than all other methods in Chapter 4.5. Incidentally, since all eigenvalues of $\mathbf{\Sigma}(\tau_{j-1}, \tau_j)$ are non-negative, the results of Theorem 4.1 also prove the following.

Corollary 4.1 *Let all the assumptions in Theorem 4.1 hold. Then as $n, p \to \infty$ such that $p/n \to c > 0$, the integrated covariance matrix estimator $\widehat{\mathbf{\Sigma}}(\tau_{j-1}, \tau_j)$ in (4.7), and also $\widehat{\mathbf{\Sigma}}(0, 1)$ in (4.8), are positive definite in probability as long as $\mathbf{\Sigma}(\tau_{j-1}, \tau_j)$ and $\mathbf{\Sigma}(0, 1)$ are.*

This corollary shows that the positive definiteness of an integrated covariance matrix is preserved in our proposed estimator in probability as we have large enough sample size. In practice, after testing different choices of n and p under simulation settings in Chapter 4.5, we always have positive definiteness of the estimator with a moderate sample size n and a similar dimension p.

Remark 4.2 *In Theorem 4.1, unlike Lam (2016), we do not require the partition to be very small with the number of data points of order smaller than the total sample size. This is because we are not proving efficiency relative to using the majority of data points in constructing the eigenmatrix for our rotation-equivariant estimator. We can pursue it, but then a very small partition essentially means $L \to \infty$ also, which unfortunately makes the rate of convergence to be slower than $n^{-1/6}$ due to the complications of microstructure noise. This can be seen explicitly in the proof of Lemma 4.4, where one of the term has rate $n^{-1/6}L$. The practical performance is also worse if we use a very small partition, resulting in too many of them. Hence we decide not to pursue something like Theorem 5 of Lam (2016), for the sake of a better rate of convergence, and a better practical performance overall.*

Remark 4.3 *The term $\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K}, v_s)\mathbf{A}(v_{s-K}, v_s)^{\mathrm{T}}\mathbf{p}_{ij}$ is bounded by $\|\mathbf{A}(v_{s-K}, v_s)\|^2$ in our proofs, defining $\mathbf{p}_{ij}$ as an eigenvector for $\mathbf{P}_{-j}$, when there are pervasive factors, which is an order p larger then when there are no factors. The same treatment goes when $\boldsymbol{\mu}_t$ is assumed non-random, where $\mathbf{A}(\cdot, \cdot)$ essentially has only one non-zero column. In the end, this is exactly the reason why $p^{3/2}/n \to c > 0$ is needed instead of just $p/n \to c > 0$. We conjecture that $p/n \to c > 0$ is enough for our results to hold even with pervasive factors, since $\mathbf{p}_{ij}$ is in fact a random eigenvector of a sample covariance-like matrix $\sum_{\ell \neq j} \widetilde{\mathbf{\Sigma}}(\tau_{j-1}, \tau_j)$. If it were a proper sample covariance matrix, then for any known unit vector $\mathbf{x} \in \mathbb{R}^p$, $\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{x} = O_P(p^{-1/2})$ (see Theorem 1 and Remark 1 of Bai et al. (2007)), so that $\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-K}, v_s)\mathbf{A}(v_{s-K}, v_s)^{\mathrm{T}}\mathbf{p}_{ij}$ should be of order $\|\mathbf{A}(v_{s-K}, v_s)\|^2/p$ in probability, i.e. the same order as when there are no factors.*

4.3.1 Extension to Jump-Diffusion Processes

Our method can be extended to include jumps in the underlying log-price process $\mathbf{X}_t$. We introduce the relevant model first. With jumps, the underlying log-price process is modeled as

$$d\mathbf{X}_t = \boldsymbol{\mu}_t dt + \boldsymbol{\sigma}_t d\mathbf{W}_t + d\mathbf{J}_t, \quad t \in [0,1], \tag{4.10}$$

where $\boldsymbol{\mu}_t$ and $\boldsymbol{\sigma}_t$ are as in the pure diffusion model (4.1), and $\mathbf{J}_t = (J_t^{(1)}, \ldots, J_t^{(p)})^{\mathrm{T}}$ denotes a p-dimensional right-continuous pure jump process. Each element in $\mathbf{J}_t$ is assumed to have finite activity in $[0,1]$, so that there are only finite number of jumps in each log-price process $X_t^{(j)}$ in the time interval we consider. The $J_t^{(j)}$'s can be correlated with each other, and each is modeled by

$$J_t^{(j)} = \sum_{\ell=1}^{N_t^{(j)}} B_\ell^{(j)}, \quad t \in [0,1],$$

where each count process $N_t^{(j)}$ can be correlated with each other. The same holds true for each jump size $B_\ell^{(j)}$. The quadratic covariation over $[0,1]$ for the process $\mathbf{X}_t$ is then

$$QV = \int_0^1 \boldsymbol{\sigma}_t \boldsymbol{\sigma}_t^{\mathrm{T}} dt + \sum_{0 \le t \le 1} \Delta\mathbf{J}_t \Delta\mathbf{J}_t^{\mathrm{T}}, \tag{4.11}$$

where $\Delta\mathbf{J}_t = \mathbf{J}_t - \mathbf{J}_{t-}$. It is clear that an off-diagonal entry in $\Delta\mathbf{J}_t\Delta\mathbf{J}_t^{\mathrm{T}}$ will only be non-zero in general when both the corresponding log-price processes have jumps at the same time (cojumps) for at least once. It can correspond to, for example, certain major market news reacted by a number of stocks at the same time. To account for the jump risks contributed by regular occurrence of cojumps (see Gilder et al. (2014) for examples of systematic or non-systematic cojumps), QV should be estimated as a whole rather than just the integrated covariance matrix.

To this end, we propose to use the wavelet method described in Section 3.2 of Fan and Wang (2007) to first remove the jumps in the log-price processes and construct our nonlinear shrinkage estimator in (4.8) using the jumps-removed data. The wavelet approach is also considered in Xue et al. (2014) to test for the presence of jumps in high-frequency financial time series. We give the practical details on how we implement the wavelet method for each observed log-price process at the end of the chapter. The estimated jump process $\widehat{\mathbf{J}}_t$ using the wavelet method is then used to construct

$\sum_{0\le t\le 1}\Delta\widehat{\mathbf{J}}_t\Delta\widehat{\mathbf{J}}_t^{\mathrm{T}}$, giving us an estimator of QV as

$$\widehat{QV}=\widehat{\mathbf{\Sigma}}(0,1)+\sum_{0\le t\le 1}\Delta\widehat{\mathbf{J}}_t\Delta\widehat{\mathbf{J}}_t^{\mathrm{T}}. \tag{4.12}$$

Note that from Theorem 1 of Fan and Wang (2007), using our notations, we can deduce immediately that the finite number of jumps in each log-price process are removed at a rate at least $(nL)^{-1/4}$ using the wavelet method, with nL being the total number of all-refresh data points. Individual asset may do even better since we use all data points available in practice for each asset before evaluating the all-refresh time points. This jump removal rate is in fact the key to the successful adaptation of wavelet jumps removal to our proposed nonlinear shrinkage estimator. More detailed assumptions:

(W1) The wavelet used in jump estimation are differentiable.

(W2) For the jump part of $X_t^{(j)}$ in $[0,1]$ for $j=1,\ldots,p$, its jump locations $\eta_\ell^{(j)}$ and jump sizes $B_\ell^{(j)}$ satisfy

$$N_1^{(j)}<\infty,\ \eta_1^{(j)}<\ldots<\eta_\ell^{(j)}<\ldots,\ 0<|B_\ell^{(j)}|<\infty \text{ almost surely.}$$

(W3) The number of stocks involved in a cojump is finite.

Assumptions (W1) and (W2) are technical assumptions adapted from Fan and Wang (2007). Assumption (W2) means that we are dealing with finite number of jumps for each log-price process, and the sizes of the jumps are bounded from 0 almost surely. If Assumption (W3) is not satisfied, then the rate of convergence of $\widehat{\mathbf{\Sigma}}(\tau j-1,\tau_j)$ in Theorem 4.1 using the jumps-removed data will be dependent on how many stocks is involved in a cojump in general. Our assumptions allow the jump process to be dependent on the drift, volatility and the microstructure noise process in general.

Theorem 4.2 *Let all the assumptions in Theorem 4.1 hold, as well as (W1) to (W3) for the jump-diffusion model (4.10). Using the jumps-removed all-refresh log-price data* $\mathbf{Y}^*(s)=\mathbf{Y}(s)-\widehat{\mathbf{J}}_{v_s}$, $s=1,\ldots,nL$ *in constructing the integrated covariance matrix estimator in (4.7), the same conclusions in Theorem 4.1 and Corollary 4.1 hold. Moreover, we have*

$$\|\sum_{0\le t\le 1}(\Delta\mathbf{J}_t\Delta\mathbf{J}_t^{\mathrm{T}}-\Delta\widehat{\mathbf{J}}_t\Delta\widehat{\mathbf{J}}_t^{\mathrm{T}})\|=O_P(n^{-1/4}).$$

The following is the jumps removal procedure:

1. Denote $Y_{i,k}^{(j)}$ the wavelet coefficients of $\{Y_t^{(j)}\}, k = 1, \ldots, 2^i,\ i = 1, \ldots, \log_2(n),\ j = 1, \ldots, p$.

2. Let $D_n^{(j)} = d\sqrt{2 \log n}$ be the universal threshold with d as the median of $|Y_{i_n,k}^{(j)}|$. If $|Y_{i_n,k}^{(j)}| > D_n^{(j)}$, the estimated jump location is $\widehat{\tau} = k2^{-i_n}$.

3. For a small neighbourhood δ_n of the estimated jump location, denote $\bar{Y}_{\widehat{\tau}_l+}^{(j)}$ and $\bar{Y}_{\widehat{\tau}_l-}^{(j)}$ as the average value over periods $[\widehat{\tau}_l, \widehat{\tau}_l + \delta_n]$ and $[\widehat{\tau}_l - \delta_n, \widehat{\tau}_l)$ respectively. We take δ_n as the square root of the total number of data points after data cleaning, following Fan and Wang (2007).

4. The estimated jump size is $\widehat{B}_l^{(j)} = \bar{Y}_{\widehat{\tau}_l+}^{(j)} - \bar{Y}_{\widehat{\tau}_l-}^{(j)}$, and the estimated jump variation is $\sum_{l=1}^{\widehat{q}} (\widehat{B}_l^{(j)})^2$, where $\widehat{q}$ is the estimated number of jumps.

5. We remove the jump effect from the original observed data as $Y_t^{*(j)} = Y_t^{(j)} - \sum_{\widehat{\tau}_l \le t} \widehat{B}_l^{(j)}$.

4.3.2 Application to Portfolio Allocation

In this chapter we investigate the theoretical performance of our estimator when it is used to construct minimum-variance portfolios. Defining $\mathbf{1}_p$ as a column vector of p ones, we define the estimated optimal minimum-variance portfolio weights to be

$$\widehat{\mathbf{w}}_{\text{opt}} = \frac{\widehat{\mathbf{\Sigma}}(0,1)^{-1}\mathbf{1}_p}{\mathbf{1}_p^{\text{T}}\widehat{\mathbf{\Sigma}}(0,1)^{-1}\mathbf{1}_p},$$

where $\widehat{\mathbf{\Sigma}}(0,1)$ is our estimator of $\mathbf{\Sigma}(0,1)$. In Chapter 4.5, we empirically compare our estimator to other estimators using different measures, including performance in minimizing portfolio risks.

Unlike DeMiguel et al. (2009) or Fan et al. (2012) which constrain the L_1 or L_2 norm of a portfolio vector $\mathbf{w}$ explicitly through a tuning parameter, our method enjoys a natural upper bound on the maximum exposure asymptotically in probability. The maximum exposure of a portfolio vector $\mathbf{w}$ is defined as $\|\mathbf{w}\|_{\max} = \max_i |w_i|$. The bound for our method is important since the theoretical minimum-variance portfolio is also subjected to the same bound. At the same time, the actual risk $R^{1/2}(\widehat{\mathbf{w}}_{\text{opt}}) = (\widehat{\mathbf{w}}_{\text{opt}}^{\text{T}}\mathbf{\Sigma}(0,1)\widehat{\mathbf{w}}_{\text{opt}})^{1/2}$ also has a natural upper bound, as presented below.

Theorem 4.3 *Let all the assumptions in Theorem 4.1 hold. Define the theoretical minimum-variance portfolio weight to be*

$$\mathbf{w}_{\text{theo}} = \frac{\mathbf{\Sigma}(0,1)^{-1}\mathbf{1}_p}{\mathbf{1}_p^{\text{T}}\mathbf{\Sigma}(0,1)^{-1}\mathbf{1}_p}.$$

In the case of no pervasive factors with $p/n \to c > 0$, or the existence of pervasive factors with $p^{3/2}/n \to c > 0$, the maximum exposures of $\widehat{\mathbf{w}}_{\text{opt}}$ and $\mathbf{w}_{\text{theo}}$ satisfy, in probability,

$$p^{1/2}\|\widehat{\mathbf{w}}_{\text{opt}}\|_{\max},\ p^{1/2}\|\mathbf{w}_{\text{theo}}\|_{\max} \leq \frac{\max_{1\leq j\leq L}\lambda_{\max}(\mathbf{\Sigma}(\tau_{j-1},\tau_j))}{\min_{1\leq j\leq L}\lambda_{\min}(\mathbf{\Sigma}(\tau_{j-1},\tau_j))},$$

where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the maximum and minimum eigenvalues of a matrix respectively.

If there are no pervasive factors and $p/n \to c > 0$, the actual risks of $\widehat{\mathbf{w}}_{\text{opt}}$ and $\mathbf{w}_{\text{theo}}$ satisfy, in probability,

$$\begin{aligned} p^{1/2}R^{1/2}(\widehat{\mathbf{w}}_{\text{opt}}) &\leq \frac{\max_{1\leq j\leq L}\lambda_{\max}(\mathbf{\Sigma}(\tau_{j-1},\tau_j))}{\min_{1\leq j\leq L}\lambda_{\min}(\mathbf{\Sigma}(\tau_{j-1},\tau_j))}\cdot\lambda_{\max}^{1/2}(\mathbf{\Sigma}(0,1)),\\ p^{1/2}R^{1/2}(\mathbf{w}_{\text{theo}}) &\leq \lambda_{\max}^{1/2}(\mathbf{\Sigma}(0,1)). \end{aligned}$$

If there are pervasive factors and $p^{3/2}n \to c > 0$, then $R(\widehat{\mathbf{w}}_{opt}) = O_P(\lambda_{\max}(\mathbf{\Sigma})) = O_P(p)$, where the bound for $R(\mathbf{w}_{theo})$ remains the same as above.

If Assumptions (W1) to (W3) hold also under the jump-diffusion model (4.10), then the same conclusions as above hold for the maximum exposure and actual risk bounds, as long as we are using the jumps-removed data as described in Chapter 4.3.1.

The proof of this theorem is in Chapter 4.7. The gross exposure constraint by Fan et al. (2012) or the L_2-norm constraint by DeMiguel et al. (2009) are useful in constraining the total exposure of a portfolio and obtaining special ones like the no-short-sale portfolio (by setting $\|\mathbf{w}\|_1 \leq 1$). In practice, as illustrated by our simulation experiments and real data analysis in Chapter 4.5, the maximum exposure can still be large while these explicit constraints are in place. Certainly, there are a lot of examples where concentrated portfolios can be rewarding. However, with respect to the minimum-variance portfolio, the theoretical one does satisfy an upper bound on the maximum exposure as presented in Theorem 4.3. Our method has the same upper

bound in probability, which decays as p increases when there are no pervasive factors in the data. As illustrated in Chapter 4.5, the maximum exposure in $\widehat{\mathbf{w}}_{\text{opt}}$ is on average smaller than other state-of-the-art methods in various settings, especially when using a quarter of a trading day as a partition. At the same time, in the real data analysis, the risk for our method measured as the out-of-sample standard deviation of the return for a portfolio is smaller than all other methods in the two portfolio studies. The relatively small turnover of our portfolio as shown in Table 4.6, 4.7, 4.8 and 4.9 is also important when profitability is concerned. See Chapter 4.5 for more details.

When there are pervasive factors like the market factor in the data, we have $\|\widehat{\mathbf{w}}_{\text{opt}}\|_{\max} = O_P(p^{1/2}) = \|\mathbf{w}_{\text{theo}}\|$, and $R(\|\widehat{\mathbf{w}}_{\text{opt}}\|) = O_P(p)$. It would seem that explicit constraints in the portfolio weights would be better than our method. However, these bounds are certainly not tight. Simulation results with pervasive factors in Tables 4.4 and 4.5 show that our method still performs better than others with the smallest L_2 distance from the theoretical portfolio, and matches closely to its out-of-sample risk. It would need more sophisticated analysis to obtain tighter bounds when there are pervasive factors.

4.4 Practical Implementation

There are two parameters that can be tuned for potentially better performance, namely the partition $(\tau_{j-1}, \tau_j]$ of the period $[0, 1]$ (thus also determining L itself which represents the number of partitions), and the scale parameter K used in the TSCV in (4.4). For example, suppose we are given a period of 10 days of tick-by-tick data, if we set $(\tau_{j-1}, \tau_j]$ to be one day, then $L = 10$. Note that the length of each partition can be different. Similar to the function $g(m)$ in equation (4.7) of Lam (2016), we propose to minimize the following criterion for a good choice of $\boldsymbol{\tau} = \{\tau_j\}_{0 \le j \le L}$ and K:

$$g(\boldsymbol{\tau}, K) = \left\| \sum_{j=1}^{L} \left(\widehat{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j) - \widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j) \right) \right\|_F^2, \tag{4.13}$$

where $\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)$ and $\widehat{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)$ are defined in (4.4) and (4.7) respectively. This function is inspired by Bickel and Levina (2008b), where a similar function, with the population covariance matrix replaced by the sample covariance matrix, is used for the determination of the banding number in banding a large covariance matrix estimator. In our case, the above aligns with the optimization problem (4.6), but with

$\mathbf{\Sigma}(\tau_{j-1}, \tau_j)$ replaced by the sample counterpart $\widetilde{\mathbf{\Sigma}}(\tau_{j-1}, \tau_j)$. From our experience, as long as the intervals are not too different in length and that each interval has enough data points (at least the same order as p), the performance of the estimator is in fact more dependent on L, the number of partitions we choose. Hence we suggest to divide the time interval into equal length partitions, checking that each one has enough data points. We can then choose L by minimizing the criterion (4.13) above. In practice, our estimator is not too sensitive to the suitable choices of L (see Table 4.2 and 4.4 for the comparison of L being one trading day (NERIVE) and L being a quarter day (quarNERIVE)).

For the choice of K, since we are using $K \asymp n^{2/3}$ as in Zhang (2011), we can search $K = [bn^{2/3}]$ on a preset grid of constant b. In practice, we found from our simulation results and real data analysis that using $b = 1$ provide good results, and portfolio performance is not too different from using other values of b, hence in this paper we use $b = 1$.

4.5 Empirical Results

4.5.1 Simulation

In this chapter, we simulate high-frequency trading transactions of 100 stocks for one year (250 trading days). The price processes and the asynchronous transaction times are simulated independently. The observed log-price is defined as $X_t^{o(i)} = X_t^{(i)} + \varepsilon_t^{(i)}$, where $X_t^{(i)}$ represents the latent log-price, and the microstructure noise has $\varepsilon_t^{(i)} \overset{iid}{\sim} N(0, 0.0005^2)$. We generate $p = 100$ latent log-prices by the following Heston-like multivariate factor model with stochastic volatilities:

$$dX_t^{(i)} = \mu^{(i)}dt + \sqrt{1-(\rho^{(i)})^2}\sigma_t^{(i)}dB_t^{(i)} + \rho^{(i)}\sigma_t^{(i)}dW_t + C\nu^{(i)}dZ_t, \quad i = 1, \ldots, 100, \tag{4.14}$$

where $\{W_t\}$, $\{Z_t\}$ and the $\{B_t^{(i)}\}$'s are independent standard Brownian motions. The processes $\{W_t\}$ and $\{Z_t\}$ imitate factors in the market. The constant $C = \mathbb{1}_{\{\text{model 2}\}}$ is 0 for the first model we consider. We set $\rho^{(i)} = -0.7C$, so that it is 0 in the first model, and hence there are no factors. For the second model, $C = 1$, so that it contains two factors. The spot volatility $\sigma_t^{(i)} = \sqrt{\varrho_t^{(i)}}$ follows the Cox-Ingersoll-Ross (CIR) process

$$d\varrho_t^{(i)} = \kappa^{(i)}(\theta^{(i)} - \varrho_t^{(i)})dt + \xi^{(i)}dU_t^{(i)},$$

where the $\{U_t^{(i)}\}$'s are independent standard Brownian motions. Other parameters of $X_t^{(i)}$ are set at $(\mu^{(i)}, \kappa^{(i)}, \xi^{(i)}, \theta^{(i)}) = (0.03x_1^{(i)}, 1.1x_2^{(i)}, 0.5x_3^{(i)}, 0.25x_4^{(i)})$ and $\nu^{(i)} = \sqrt{\theta^{(i)}}$, where the $x_j^{(i)}$'s are independent uniform random variables on the interval $[0.7, 1.3]$. The initial value of each log-price $X_0^{(i)}$ is set randomly on the interval $[0.5, 1.5]$ and the starting spot volatility $\varrho_0^{(i)}$ on the interval $[0.5\theta^{(i)}, 1.5\theta^{(i)}]$.

For the transaction times, we generate 100 different Poisson processes with intensities $\lambda_1, \ldots, \lambda_{100}$ respectively. Since the normal trading time for one day is 23400 seconds, λ_i is set to be $0.01i \times 23400$, where $i = 1, \ldots, 100$.

4.5.2 Comparison of Different Estimators

Comparisons with TSCV and thresholded method

We compare out estimator to the TSCV, as well as the Thresholded Average Realized Volatility Matrix (TARVM) which is essentially a thresholded TSCV introduced in Wang and Zou (2010). The reason we choose to compare to the TARVM on top of the TSCV is because when there are no factors, sparseness or approximate sparseness in $\mathbf{\Sigma}(0,1)$ can be natural as its eigenvalues are of constant order even with a diverging matrix dimension, giving potential advantages to thresholded estimators. Our estimator is a modified TSCV, and so comparing to another modified TSCV like the TARVM makes sense. Hereafter, we abbreviate our estimator as NERIVE when we are using one trading day as a partition length, and quarNERIVE when we are using a quarter of a trading day.

We use two measures for comparing the estimators. One is the Frobenius error, another is the average bias in eigenvalues, defined by

$$\text{Frobenius error} = \text{tr}(\widehat{\mathbf{\Sigma}}(0,1) - \mathbf{\Sigma}(0,1))^2, \quad \text{Average bias} = \text{tr}(\widehat{\mathbf{\Sigma}}(0,1) - \mathbf{\Sigma}(0,1))/p.$$

The integrated covariance matrix $\mathbf{\Sigma}(0,1)$ is evaluated using the simulated latent log-prices at the finest grid (1 per second). We divide the 250 trading days into disjoint 5-day intervals, and calculate the two error measures for different estimators over each 5-day interval. The means and standard deviations of these errors are reported in Table 4.1. It also includes the same exercise when 5-day becomes 1-day intervals.

When we are using 5-day training windows, with roughly 200 points per day after all refresh method, it is clear that NERIVE, especially quarNERIVE, performs better

No factors ($C = 0$)		NERIVE	quarNERIVE	TSCV	TARVM
5-day	Frobenius error	$12_{(1.3)}$	$7_{(0.9)}$	$156_{(22.2)}$	$64_{(5.3)}$
	Average bias	$30_{(1.6)}$	$23_{(1.4)}$	$33_{(2.3)}$	$36_{(1.6)}$
1-day	Frobenius error	-	$0.4_{(0.1)}$	$9.3_{(1.7)}$	$1.9_{(0.2)}$
	Average bias	-	$3_{(0.4)}$	$2_{(0.5)}$	$1_{(0.2)}$
With factors ($C = 1$)		NERIVE	quarNERIVE	TSCV	TARVM
5-day	Frobenius error	$2007_{(1269)}$	$1161_{(539)}$	$3241_{(3370)}$	$3123_{(1527)}$
	Average bias	$59_{(14)}$	$45_{(8)}$	$67_{(25)}$	$72_{(14)}$
1-day	Frobenius error	-	$40_{(32)}$	$62_{(51)}$	$12_{(11)}$
	Average bias	-	$7_{(3.7)}$	$4_{(5.9)}$	$3_{(2.8)}$

Table 4.1 Mean and standard deviation of Frobenius error and average bias of eigenvalues over different 5-day or 1-day intervals for various methods. All values are multiplied by 10000.

than TSCV and TAVRM in both measures. However, in using 1-day training windows, TAVRM is better in terms of average bias in the eigenvalues. When there are factors, TAVRM is also better in Frobenius norm error using 1-day training windows. It is clear that there are advantages in thresholding, especially when we consider a shorter window for the integrated covariance matrix, but our method is better in general when such window increases.

Comparisons with POET and related methods

POET, originally proposed as a general low-frequency data method in Fan et al. (2013), essentially assumes that the true covariance matrix can be decomposed into a low rank matrix (induced from factors in the data) plus a sparse residual one. Aït-Sahalia and Xiu (2017) proposes such a decomposition on the realized covariance matrix of sub-sampled return data (15 or 30 minutes interval) to reduce the effects of microstructure noise contamination, while the residual covariance is assumed to be block diagonal with known blocks (e.g. blocking by industry). Dai et al. (2017) proposed the POET method on realized covariance matrix calculated on pre-averaged return data (PRVM), with thresholding developed for the residual matrix. We find that such thresholding usually works better than blocking using industry, and so we compare our method to such a POET method applied on TSCV (TS-POET), since NERIVE or quarNERIVE are based on nonlinear shrinkage of the TSCV.

We also explore if our nonlinear shrinkage can be applied to the PRVM rather than TSCV. To be precise, we replace $\widetilde{\mathbf{\Sigma}}(\tau_{\ell-1}, \tau_\ell)$ in (4.4) by the corresponding PRVM, and follow Chapter 4.2.2 to construct $\widehat{\mathbf{\Sigma}}(0,1)$ in (4.8). We abbreviate nonlinear shrinkage based on PRVM as PR-NERIVE or PR-quarNERIVE, and compare them with the POET in Dai et al. (2017) (PR-POET). Since Fan and Kim (2017) has developed a robust version of PRVM with POET (RPR-POET), we compare this to PR-NERIVE and PR-quarNERIVE as well. Throughout the rest of the chapter, all POET methods use 5 factors which is enough to achieve consistently good results.

Figure 4.1 shows the Frobenius errors when there are no factors in model (4.14). It is clear that quarNERIVE is better than NERIVE and TS-POET. When we use pre-averaged data for nonlinear shrinkage, PR-quarNERIVE is also better than PR-NERIVE, PR-POET and RPR-POET. When there are factors in model (4.14), quarNERIVE is still better than NERIVE and TS-POET, but PR-quarNERIVE is not as good as RPR-POET when we consider a longer time horizon for the integrated

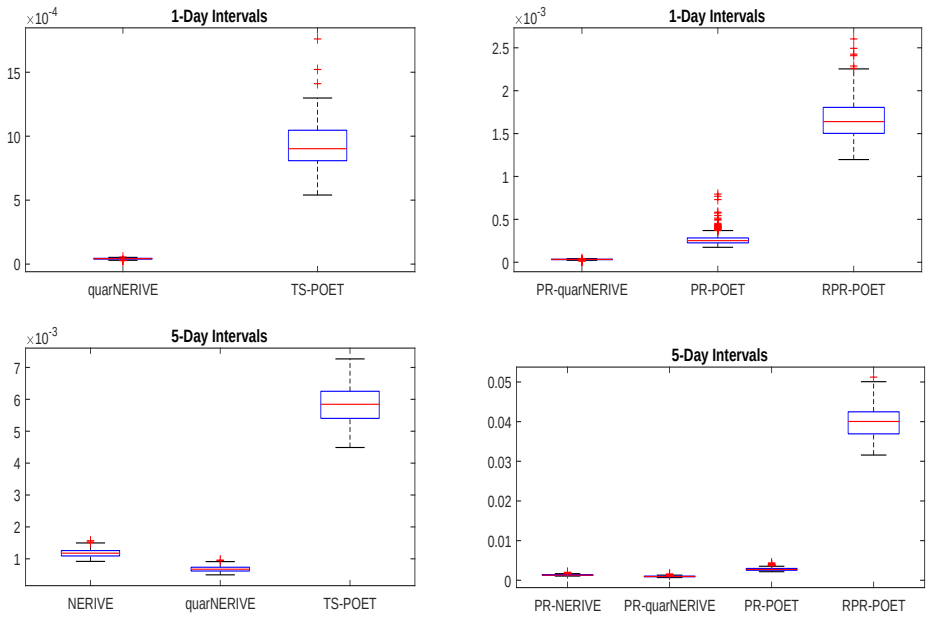


Fig. 4.1 Boxplot of Frobenius errors when there are no factors in model (4.14) ($C = 0$).

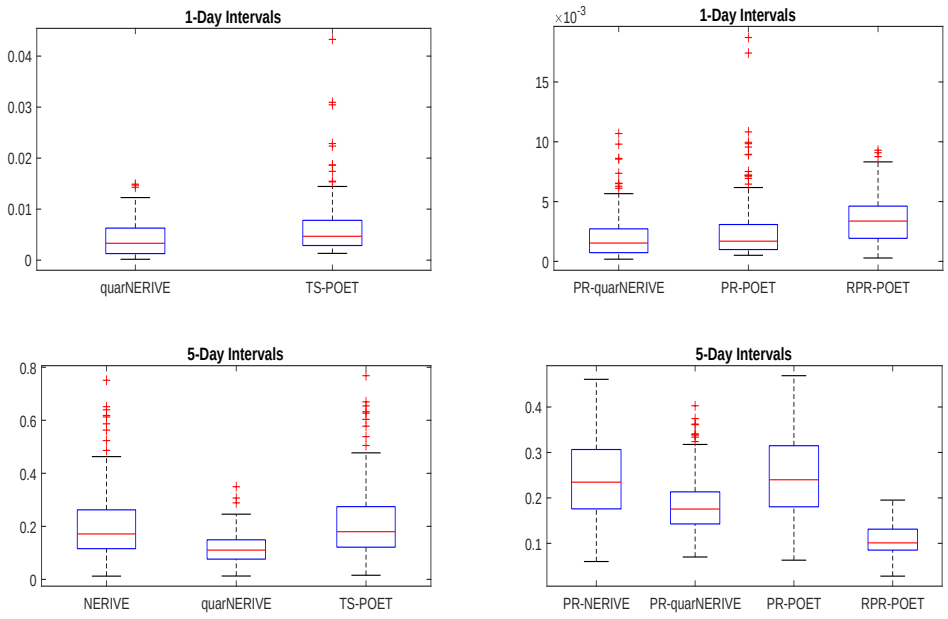


Fig. 4.2 Boxplot of Frobenius errors when there are factors in model (4.14) ($C = 1$).

covariance matrix (5-day) in Figure 4.2. Clearly PR-NERIVE has a lot of potential, and we hope to develop its theoretical performances in another project (see Remark 4.1 as well). We have also considered the spectral error, but the patterns are very similar to Figures 4.1 and 4.2, and hence they are omitted.

4.5.3 Comparison of Portfolio Allocation Performance

To compare the performance of different methods, we focus on the minimum-variance portfolio

$$\mathbf{w}_{\text{opt}} = \frac{\widehat{\mathbf{\Sigma}}(0,1)^{-1}\mathbf{1}_p}{\mathbf{1}_p^{\mathrm{T}}\widehat{\mathbf{\Sigma}}(0,1)^{-1}\mathbf{1}_p}, \quad \text{which solves} \quad \min_{\mathbf{w}:\mathbf{w}^{\mathrm{T}}\mathbf{1}_p=1} \mathbf{w}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}(0,1)\mathbf{w}.$$

We first set the benchmark for comparisons. Following Fan et al. (2012), we create a theoretical portfolio $\mathbf{w}_{\text{theo}}$, which is a minimum variance portfolio with $\mathbf{\Sigma}(0,1)$ evaluated similarly as in Chapter 4.5.2. For all other methods, we use the all-refresh time points evaluated from the data (we do not hold positions overnight for all methods to avoid overnight price jumps, since they are not what our study is about).

Other portfolios are constructed and compared to the theoretical minimum variance portfolio (THEO) above. The first one is the equal weight portfolio (EQUAL). The second one is the minimum variance portfolio with $\mathbf{\Sigma}(0,1)$ substituted by the Two Scale CoVariance matrix (TSCV). We abbreviate it as TARVM when $\mathbf{\Sigma}(0,1)$ is replaced by the TARVM as in Chapter 4.5.2. When $\mathbf{\Sigma}(0,1)$ is substituted with our estimator, we abbreviate it as NERIVE with one trading day as a partition length, and quarNERIVE when a partition length is a quarter of a trading day. We also compare with the Gross Exposure Constraint (GEC) method (Fan et al., 2012), and finally the L_2 norm constraint (NORM) (DeMiguel et al., 2009) based on TSCV. The GEC and NORM methods solve respectively

$$\text{GEC:} \min_{\mathbf{w}:\mathbf{w}^{\mathrm{T}}=1,\|\mathbf{w}\|_1\le c} \mathbf{w}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}(0,1)\mathbf{w},$$
$$\text{NORM:} \min_{\mathbf{w}:\mathbf{w}^{\mathrm{T}}=1,\|\mathbf{w}\|_2^2\le\delta} \mathbf{w}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}(0,1)\mathbf{w}.$$

We constructed 3 GEC portfolios with tuning parameters $c = 1, 2, 3$, as well as 3 NORM portfolios with tuning parameters $\delta = 0.1, 0.5, 1$ for comparisons. We do not use the pairwise refresh method for GEC to save significant computational time in

both the simulations and the real data analysis, as well as that the features of our method can be compared more directly to those of GEC. Finally, we also compare to TS-POET, PR-NERIVE, PR-quarNERIVE, PR-POET and RPR-POET when the corresponding estimator substitutes $\mathbf{\Sigma}(0,1)$ in $\mathbf{w}_{\text{opt}}$.

The portfolio exercise is carried out as follows for all methods. We invest 1 unit of capital to the different portfolios above at a certain start date (e.g., day 6 if we are using a 5-day training window), and rebalance the portfolio weights daily, moving the training window one day ahead. There are two investment strategies for comparisons under each model 1 or 2. The first one rebalances the portfolio daily with a 5-day training window. The second one rebalances the portfolio daily with a 1-day training window.

The quantities to be compared for different portfolios are as follows. For daily rebalancing with a k-day training window ($k = 1$ or 5), we calculate the annualized portfolio return and annualized out-of-sample standard deviation, given respectively by

$$\hat{\mu} = 250 \times \frac{1}{250-k} \sum_{i=k+1}^{250} \mathbf{w}^{\mathrm{T}} \mathbf{r}_i, \quad \hat{\sigma} = \left(250 \times \frac{1}{250-k} \sum_{i=k+1}^{250} (\mathbf{w}^{\mathrm{T}} \mathbf{r}_i - \frac{\hat{\mu}}{250})^2 \right)^{1/2}.$$

The out-of-sample standard deviation is a good indicator of how much risk is associated with a portfolio (DeMiguel et al., 2009), and is our main quantity for performance comparisons, whereas portfolio return is of secondary importance. We also calculate the Sharpe ratio $\hat{\mu}/\hat{\sigma}$. The average maximum exposure and the maximum of the maximum exposure over the whole investment horizon are two important measures for comparisons too. Since this is a simulation experiment, we can calculate the actual risk of a portfolio $\mathbf{w}$, $R^{1/2}(\mathbf{w}) = (\mathbf{w}^{\mathrm{T}} \mathbf{\Sigma} \mathbf{w})^{1/2}$, over a trading day. We compare the averaged actual risks of different methods over the whole investment horizon. Finally we compare the error norm compared to $\mathbf{w}_{\text{theo}}$, defined as Norm $= \|\mathbf{w} - \mathbf{w}_{\text{theo}}\|$, and also the portfolio turnover for different methods.

Table 4.2 and 4.3 show the results for model (4.14) with no factors. Excluding all methods based on pre-averaged return data, the out-of-sample standard deviations of NERIVE and quarNERIVE are among the smallest for both 5-day and 1-day training windows, and closely match that of the theoretical minimum portfolio. TS-POET is the best when we are using 1-day training window. Pre-averaging tends to improve on nonlinear shrinkage and POET also, with PR-POET the best when we are using 1-day training window. The equal weight portfolio performs well also but is not as good as

our methods when we use 5-day training windows. Our methods also have (together with TS-POET and PR-POET) the closest L_2 distance from the theoretical minimum portfolio weight, and apart from GEC1, PR-quarNERIVE has the smallest portfolio turnover. Both TSCV and TARVM are having much larger actual risks than other methods, and a lot of times with impractical maximum exposures.

Table 4.4 and 4.5 show the results for model (4.14) with factors. In general, risks are higher with factors, even for the theoretical portfolio. Our methods (quarNERIVE or PR-quarNERIVE) have risks close to the theoretical ones, with portfolio weights the closest to the theoretical portfolio weights among all methods. Equal weight portfolio now performs at a similar level to other methods (apart from TSCV and TARVM) in terms of risk minimization, but our methods are around 50% better in minimizing the our-of-sample SD or the actual risk. TSCV and TARVM are still the worst in terms of risks, maximum exposures and portfolio turnover. Overall, NERIVE or quarNERIVE (and their pre-averaging versions) do well in risk minimization compared to all other methods including the equal weight portfolio, with reasonable and often small maximum exposures and portfolio turnover.

4.5.4 Portfolio Allocation Study

In this study, we choose the stocks based on two lists, the "Fifty Most Active Stocks on NYSE, Round Lots (mils. of shares), 2013" and "Fifty Most Active Stocks by Dollar Volume on NYSE ($ in mils.), 2013", from the New York Stock Exchange Data official website `http://www.nyxdata.com/`. There are 26 stocks appearing in both of the lists above, and 74 stocks in either of them. We downloaded all the trading transactions of these 74 stocks in Year 2013 from Wharton Research Data Services (WRDS, `https://wrds-web.wharton.upenn.edu/`). We omit the stock Sprint Corporation due to missing price data. We first clean all the data by the R-package "highfrequency", which follows the high-frequency data cleaning steps presented in Barndorff-Nielsen et al. (2009). We conduct our portfolio allocation study on two portfolios, one with the $p = 26$ stocks appearing in both lists, and the other with $p = 73$ stocks appearing in either of the lists.

We carry out the same portfolio allocation exercises as in our simulations for both the 26-stock and 73-stock portfolios. First we do not remove jumps from the cleaned data. The results are displayed in Tables 4.6, 4.7, 4.8 and 4.9. Both NERIVE and quarNERIVE achieve the lowest out-of-sample SD in the two scenarios presented for

Methods	Out-of-Sample SD (%)	Actual Risk(%)	Norm	Aver Max Abs Weight(%)	Max Max Abs Weight(%)	Portfolio Turnover	Portfolio Return(%)	Sharpe Ratio
daily rebalancing portfolio with 5-day training window								
THEO	1.6	1.7	–	$10_{(6.2)}$	44	$0.06_{(0.02)}$	5.2	3.2
NERIVE	1.9	1.9	0.08	$6_{(2.7)}$	18	$0.14_{(0.02)}$	7.4	3.9
quarNERIVE	1.8	1.9	0.07	$7_{(3.0)}$	19	$0.12_{(0.02)}$	6.0	3.3
EQUAL	2.0	2.2	0.13	$1_{(-)}$	1	–	5.5	2.8
TSCV	149.6	149.2	1.34	$64_{(352.7)}$	5066	$4.21_{(33.43)}$	297.5	2.0
GEC1	2.1	2.3	0.13	$2_{(1.0)}$	7	$0.06_{(0.04)}$	6.1	2.9
GEC2	2.5	2.6	0.13	$8_{(4.0)}$	32	$0.35_{(0.06)}$	3.9	1.6
GEC3	2.9	2.9	0.15	$7_{(2.1)}$	14	$0.42_{(0.08)}$	4.5	1.5
NORM0.1	3.6	3.5	0.18	$8_{(3.6)}$	24	$0.62_{(0.24)}$	9.8	2.7
NORM0.5	5.8	5.3	0.28	$16_{(8.9)}$	54	$1.20_{(0.51)}$	2.0	0.3
NORM1	7.1	6.6	0.36	$20_{(13.8)}$	80	$1.56_{(0.92)}$	7.9	1.1
TARVM	7.9	14.5	0.44	$34_{(119.6)}$	1276	$1.47_{(4.27)}$	−21.6	−2.7
TS-POET	2.0	2.0	0.08	$9_{(4.1)}$	27	$0.15_{(0.03)}$	9.7	4.9
PR-NERIVE	1.8	1.9	0.08	$6_{(2.6)}$	14	$0.11_{(0.02)}$	6.6	3.6
PR-quarNERIVE	1.8	1.9	0.08	$6_{(2.5)}$	13	$0.10_{(0.02)}$	6.0	3.3
PR-POET	1.8	1.9	0.06	$9_{(4.3)}$	27	$0.10_{(0.02)}$	6.0	3.4
RPR-POET	5.6	6.0	0.32	$26_{(7.3)}$	52	$0.36_{(0.15)}$	1.0	0.2

Table 4.2 Simulation results for model 1 with 5-day training window and no factors ($C = 0$ in (4.14))

* Annualized out-of-sample standard deviation, actual risk, norm of weights difference, averaged maximum absolute weight (standard deviation in bracket), maximum of maximum absolute weight, portfolio return and Sharpe ratio for various methods, including GEC ($c = 1, 2, 3$) and NORM ($\delta = 0.1, 0.5, 1$).

Methods	Out-of-Sample SD (%)	Actual Risk(%)	Norm	Aver Max Abs Weight(%)	Max Max Abs Weight(%)	Portfolio Turnover	Portfolio Return(%)	Sharpe Ratio
daily rebalancing portfolio with 1-day training window								
THEO	1.6	1.7	–	$10_{(6.2)}$	44	$0.06_{(0.02)}$	5.7	3.6
quarNERIVE	2.1	2.2	0.13	$2_{(0.3)}$	4	$0.2_{(0.03)}$	5.4	2.6
EQUAL	2.0	2.2	0.13	$1_{(-)}$	1	–	6.1	3.1
TSCV	373.4	806.1	5.43	$172_{(1187.5)}$	17696	$15.95_{(117.09)}$	1204.9	3.2
GEC1	2.0	2.2	0.13	$1_{(0.1)}$	2	$0.04_{(0.01)}$	6.0	3.0
GEC2	6.9	7.1	0.29	$11_{(6.6)}$	36	$1.02_{(0.32)}$	−5.0	−0.7
GEC3	16.5	16.1	0.63	$28_{(11.1)}$	58	$2.38_{(1.10)}$	−45.6	−2.8
NORM0.1	6.6	6.5	0.28	$7_{(2.8)}$	13	$1.28_{(0.31)}$	2.8	0.4
NORM0.5	14.8	14.7	0.60	$19_{(6.0)}$	32	$3.38_{(1.14)}$	−0.5	0.0
NORM1	20.7	19.3	0.80	$25_{(8.1)}$	51	$-12.41_{(268.64)}$	2.5	0.1
TARVM	342.8	511.0	4.64	$139_{(652.4)}$	7949	$14.63_{(144.97)}$	902.8	2.6
TS-POET	1.8	2.0	0.08	$9_{(4.4)}$	28	$0.29_{(0.03)}$	3.3	1.8
PR-quarNERIVE	2.0	2.2	0.12	$2_{(0.3)}$	3	$0.17_{(0.02)}$	5.3	2.6
PR-POET	1.7	1.9	0.06	$9_{(5.3)}$	35	$0.21_{(0.03)}$	5.7	3.3
RPR-POET	5.8	6.2	0.31	$25_{(9.1)}$	66	$0.98_{(0.32)}$	−1.4	−0.2

Table 4.3 Simulation results for model 1 with 1-day training window and no factors ($C = 0$ in (4.14))

* Annualized out-of-sample standard deviation, actual risk, norm of weights difference, averaged maximum absolute weight (standard deviation in bracket), maximum of maximum absolute weight, portfolio return and Sharpe ratio for various methods, including GEC ($c = 1, 2, 3$) and NORM ($\delta = 0.1, 0.5, 1$).

Methods	Out-of-Sample SD (%)	Actual Risk(%)	Norm	Aver Max Abs Weight(%)	Max Max Abs Weight(%)	Portfolio Turnover	Portfolio Return(%)	Sharpe Ratio
daily rebalancing portfolio with 5-day training window								
THEO	13	13	–	$41_{(21)}$	143	$0.3_{(0.1)}$	−1.5	−0.1
NERIVE	14	15	0.55	$22_{(5)}$	42	$0.9_{(0.2)}$	0.8	0.1
quarNERIVE	14	14	0.53	$23_{(6)}$	48	$0.8_{(0.2)}$	−10.7	−0.8
EQUAL	27	27	0.97	$1_{(-)}$	1	–	28.8	1.1
TSCV	34235	19343	107.02	$3937_{(51415)}$	804861	$468.0_{(6900.5)}$	−75911.4	−2.2
GEC1	25	25	1.00	$45_{(15)}$	94	$0.5_{(0.2)}$	20.2	0.8
GEC2	24	24	1.00	$47_{(16)}$	98	$0.7_{(0.3)}$	20.3	0.9
GEC3	23	24	1.01	$47_{(15)}$	106	$0.9_{(0.5)}$	14.8	0.6
NORM0.1	21	22	0.87	$8_{(1)}$	14	$0.6_{(0.2)}$	5.5	0.3
NORM0.5	17	18	0.78	$18_{(2)}$	26	$1.2_{(0.4)}$	−44.7	−2.6
NORM1	18	18	0.86	$28_{(3)}$	44	$1.7_{(0.6)}$	−49.2	−2.8
TARVM	8777	13790	91.58	$3077_{(33547)}$	524587	$61.7_{(447.5)}$	18387.1	2.1
TS-POET	17	17	0.77	$31_{(7)}$	59	$1.8_{(0.3)}$	−5.9	−0.4
PR-NERIVE	14	14	0.53	$21_{(5)}$	42	$0.9_{(0.1)}$	−4.7	−0.3
PR-quarNERIVE	14	14	0.52	$22_{(5)}$	39	$0.8_{(0.1)}$	−11.5	−0.8
PR-POET	15	15	0.57	$32_{(8)}$	60	$1.2_{(0.2)}$	−11.2	−0.7
RPR-POET	19	20	0.96	$41_{(10)}$	81	$2.4_{(0.7)}$	−16.9	−0.9

Table 4.4 Simulation results for model 2 with 5-day training window and factors ($C = 1$ in (4.14))

* Annualized out-of-sample standard deviation, actual risk, norm of weights difference, averaged maximum absolute weight (standard deviation in bracket), maximum of maximum absolute weight, portfolio return and Sharpe ratio for various methods, including GEC ($c = 1, 2, 3$) and NORM ($\delta = 0.1, 0.5, 1$).

Methods	Out-of-Sample SD (%)	Actual Risk(%)	Norm	Aver Max Abs Weight(%)	Max Max Abs Weight(%)	Portfolio Turnover	Portfolio Return(%)	Sharpe Ratio
daily rebalancing portfolio with 1-day training window								
THEO	13	13	–	$41_{(21)}$	143	$0.3_{(0.1)}$	0.4	0.0
quarNERIVE	16	17	0.75	$17_{(2)}$	26	$2.0_{(0.6)}$	−37.5	−2.3
EQUAL	27	27	0.97	$1_{(-)}$	1	–	20.8	0.8
TSCV	1605	2061	30.52	$1054_{(5474)}$	77945	$39.7_{(439.9)}$	112.1	0.1
GEC1	26	27	1.05	$44_{(17)}$	95	$0.8_{(0.2)}$	−6.0	−0.2
GEC2	26	26	1.05	$46_{(16)}$	103	$1.1_{(0.3)}$	−12.8	−0.5
GEC3	25	25	1.06	$48_{(16)}$	106	$1.6_{(0.6)}$	−27.1	−1.1
NORM0.1	22	23	0.89	$8_{(1)}$	14	$1.4_{(0.2)}$	4.2	0.2
NORM0.5	21	22	0.95	$19_{(3)}$	33	$3.1_{(1.0)}$	3.4	0.2
NORM1	26	26	1.23	$29_{(5)}$	52	$5.7_{(5.3)}$	22.2	0.8
TARVM	2707	2714	23.87	$667_{(4307)}$	67297	$1.3_{(536.7)}$	5567.6	2.1
TS-POET	20	20	0.84	$22_{(6)}$	51	$2.5_{(0.3)}$	−1.1	−0.1
PR-quarNERIVE	16	17	0.72	$16_{(2)}$	25	$2.0_{(0.2)}$	−39.9	−2.5
PR-POET	17	18	0.76	$27_{(8)}$	56	$2.7_{(0.3)}$	11.7	0.7
RPR-POET	23	24	1.21	$51_{(22)}$	159	$5.1_{(2.2)}$	64.7	2.8

Table 4.5 Simulation results for model 2 with 1-day training window and factors ($C = 1$ in (4.14))

* Annualized out-of-sample standard deviation, actual risk, norm of weights difference, averaged maximum absolute weight (standard deviation in bracket), maximum of maximum absolute weight, portfolio return and Sharpe ratio for various methods, including GEC ($c = 1, 2, 3$) and NORM ($\delta = 0.1, 0.5, 1$).

$\mathbf{p = 26}$ Methods	Out-of-Sample SD (%)	Aver Max Abs Weight(%)	Max Max Abs Weight(%)	Portfolio Turnover	Portfolio Return(%)	Sharpe Ratio
daily rebalancing portfolio with 5-day training window						
NERIVE	4.5	$21_{(6)}$	41	$0.26_{(0.1)}$	18.6	4.2
quarNERIVE	4.4	$20_{(5)}$	36	$0.22_{(0.1)}$	21.0	4.8
EQUAL	5.2	$4_{(-)}$	4	–	24.3	4.6
TSCV	6.1	$42_{(13)}$	84	$1.16_{(0.5)}$	16.9	2.8
GEC1	5.0	$30_{(11)}$	69	$0.33_{(0.1)}$	28.0	5.6
GEC2	4.9	$34_{(11)}$	78	$0.58_{(0.2)}$	20.1	4.1
GEC3	5.4	$39_{(12)}$	83	$0.88_{(0.3)}$	18.0	3.3
NORM0.1	4.6	$13_{(2)}$	19	$0.25_{(0.1)}$	18.6	4.0
NORM0.5	5.4	$33_{(6)}$	52	$0.84_{(0.3)}$	14.3	2.7
NORM1	5.9	$41_{(11)}$	74	$1.09_{(0.4)}$	14.7	2.5
TARVM	13.3	$56_{(75)}$	1097	$1.74_{(3.0)}$	−9.1	−0.7
TS-POET	5.0	$30_{(8)}$	59	$0.60_{(0.2)}$	18.9	3.7
PR-NERIVE	4.4	$19_{(5)}$	39	$0.23_{(0.1)}$	18.1	4.1
PR-quarNERIVE	4.3	$19_{(5)}$	36	$0.22_{(0.1)}$	19.4	4.5
PR-POET	4.4	$25_{(7)}$	48	$0.34_{(0.1)}$	18.2	4.1
RPR-POET	8.0	$50_{(15)}$	110	$1.12_{(0.4)}$	21.0	2.6

Table 4.6 Empirical results for the 26 most actively traded stocks in NYSE: 5-day training window

* Annualized out-of-sample standard deviation, averaged maximum absolute weight, maximum of maximum absolute weight, annualized portfolio return and Sharpe ratio for various methods, including GEC ($c = 1, 2, 3$) and NORM ($\delta = 0.1, 0.5, 1$).

p = 26 Methods	Out-of-Sample SD (%)	Aver Max Abs Weight(%)	Max Max Abs Weight(%)	Portfolio Turnover	Portfolio Return(%)	Sharpe Ratio
daily rebalancing portfolio with 1-day training window						
quarNERIVE	4.5	$19_{(7)}$	42	$0.58_{(0.2)}$	17.8	3.9
EQUAL	5.2	$4_{(-)}$	4	–	24.1	4.6
TSCV	7.0	$60_{(80)}$	973	$3.15_{(7.3)}$	19.6	2.8
GEC1	4.9	$28_{(13)}$	67	$0.54_{(0.2)}$	28.2	5.7
GEC2	5.0	$33_{(12)}$	74	$0.97_{(0.2)}$	26.7	5.3
GEC3	5.5	$36_{(13)}$	87	$1.32_{(0.3)}$	24.5	4.4
NORM0.1	4.6	$12_{(2)}$	19	$0.52_{(0.1)}$	20.7	4.5
NORM0.5	5.3	$29_{(6)}$	47	$1.33_{(0.2)}$	18.3	3.5
NORM1	5.8	$41_{(10)}$	70	$1.86_{(0.4)}$	17.2	3.0
TARVM	13.4	$40_{(93)}$	878	$3.49_{(18.5)}$	3.4	0.3
TS-POET	4.8	$28_{(11)}$	85	$1.24_{(0.4)}$	22.7	4.7
PR-quarNERIVE	4.5	$18_{(6)}$	38	$0.55_{(0.2)}$	16.5	3.7
PR-POET	4.8	$26_{(9)}$	59	$1.00_{(0.3)}$	22.1	4.6
RPR-POET	9.2	$67_{(27)}$	193	$2.50_{(0.6)}$	−2.2	−0.2

Table 4.7 Empirical results for the 26 most actively traded stocks in NYSE: 1-day training window

* Annualized out-of-sample standard deviation, averaged maximum absolute weight, maximum of maximum absolute weight, annualized portfolio return and Sharpe ratio for various methods, including GEC ($c = 1, 2, 3$) and NORM ($\delta = 0.1, 0.5, 1$).

p = 73 Methods	Out-of-Sample SD (%)	Aver Max Abs Weight(%)	Max Max Abs Weight(%)	Portfolio Turnover	Portfolio Return(%)	Sharpe Ratio
daily rebalancing portfolio with 5-day training window						
NERIVE	3.8	$12_{(3)}$	22	$0.44_{(0.1)}$	15.3	4.0
quarNERIVE	3.9	$12_{(3)}$	21	$0.40_{(0.1)}$	16.1	4.1
EQUAL	5.4	$1_{(-)}$	1	–	22.3	4.1
TSCV	470.8	$629_{(4042)}$	58950	$104.71_{(1554.4)}$	1367.1	2.9
GEC1	5.0	$21_{(11)}$	57	$0.34_{(0.2)}$	21.4	4.3
GEC2	4.7	$25_{(10)}$	64	$0.57_{(0.2)}$	14.2	3.0
GEC3	4.7	$25_{(9)}$	59	$0.87_{(0.2)}$	9.5	2.0
NORM0.1	4.5	$9_{(1)}$	15	$0.46_{(0.1)}$	14.7	3.3
NORM0.5	4.9	$19_{(5)}$	33	$1.29_{(0.3)}$	8.8	1.8
NORM1	5.8	$26_{(7)}$	48	$2.07_{(0.6)}$	7.9	1.4
TARVM	4.6	$5_{(1)}$	12	$0.09_{(0.0)}$	21.5	4.7
TS-POET	4.2	$19_{(5)}$	37	$0.87_{(0.3)}$	14.6	3.4
PR-NERIVE	3.9	$11_{(3)}$	21	$0.43_{(0.1)}$	15.7	4.0
PR-quarNERIVE	3.9	$11_{(3)}$	22	$0.39_{(0.1)}$	15.3	3.9
PR-POET	3.8	$16_{(5)}$	32	$0.56_{(0.2)}$	17.0	4.4
RPR-POET	5.6	$15_{(4)}$	33	$0.61_{(0.2)}$	22.6	4.0

Table 4.8 Empirical results for the 73 most actively traded stocks in NYSE: 5-day training window

* Annualized out-of-sample standard deviation, averaged maximum absolute weight, maximum of maximum absolute weight, annualized portfolio return and Sharpe ratio for various methods, including GEC ($c = 1, 2, 3$) and NORM ($\delta = 0.1, 0.5, 1$).

p = 73 Methods	Out-of-Sample SD (%)	Aver Max Abs Weight(%)	Max Max Abs Weight(%)	Portfolio Turnover	Portfolio Return(%)	Sharpe Ratio
daily rebalancing portfolio with 1-day training window						
quarNERIVE	4.2	$8_{(3)}$	20	$0.79_{(0.2)}$	18.6	4.4
EQUAL	5.4	$1_{(-)}$	1	–	22.4	4.2
TSCV	120.3	$381_{(671)}$	3897	$25.80_{(214.4)}$	−33.2	−0.3
GEC1	5.1	$7_{(13)}$	69	$0.17_{(0.2)}$	19.2	3.7
GEC2	5.5	$19_{(12)}$	61	$0.83_{(0.2)}$	25.3	4.6
GEC3	9.7	$30_{(27)}$	158	$1.58_{(3.9)}$	29.8	3.1
NORM0.1	4.6	$8_{(3)}$	19	$0.76_{(0.2)}$	21.4	4.6
NORM0.5	8.0	$18_{(10)}$	50	$1.95_{(0.9)}$	29.7	3.7
NORM1	11.0	$27_{(17)}$	69	$3.25_{(3.5)}$	21.4	1.9
TARVM	103.9	$221_{(455)}$	4726	$44.63_{(324.1)}$	136.7	1.3
TS-POET	5.2	$23_{(11)}$	83	$1.78_{(0.5)}$	15.8	3.0
PR-quarNERIVE	4.3	$8_{(2)}$	18	$0.79_{(0.2)}$	16.9	3.9
PR-POET	4.2	$13_{(5)}$	40	$0.94_{(0.3)}$	20.1	4.8
RPR-POET	6.4	$17_{(5)}$	43	$1.10_{(0.2)}$	23.7	3.7

Table 4.9 Empirical results for the 73 most actively traded stocks in NYSE: 1-day training window

* Annualized out-of-sample standard deviation, averaged maximum absolute weight, maximum of maximum absolute weight, annualized portfolio return and Sharpe ratio for various methods, including GEC ($c = 1, 2, 3$) and NORM ($\delta = 0.1, 0.5, 1$).

both portfolios, which are all under 4.5%. PR-POET has similarly good performance too, although with higher maximum exposure compared to NERIVE, quarNERIVE and their pre-averaging versions. The maximum exposure of our methods are reasonably low even compared to the no short sale or L_2-constraint portfolios, with among the lowest portfolio turnovers under all scenarios for both $p = 26$ and $p = 73$ portfolios.

We also considered jumps removed data. The results are presented in Table 4.10, Table 4.11, Table 4.12 and Table 4.13. In general, the out-of-sample SD do not change much for all methods, except for TSCV and TARVM which can see huge increase or decrease in the risk. It is not surprising though as both methods can invest heavily in all stocks, rendering them more sensitive to jumps removal. In fact the number of jumps estimated for each date is typically around 4 or 5, which is a very small number compared to the number of all-refresh data points.

4.6 Conclusion

We generalize nonlinear shrinkage of eigenvalues in a large sample covariance matrix for independent and identically distributed random vectors (Lam, 2016) to that of a large two-scale covariance matrix estimator (TSCV) for high-frequency returns, which are not independent and identically distributed in general. To do this, we split the data into partitions and regularize the eigenvalues of the TSCV within a partition by the data from other partitions. Regularization is indeed achieved both theoretically and empirically, as demonstrated by the good performance in our simulations and portfolio allocation exercises.

Since TSCV has a slower rate of convergence than the multi-scale realized volatility matrix (Tao et al., 2013), the kernel realized volatility matrix (Barndorff-Nielsen et al., 2011) or the pre-averaging realized volatility positive semi-definite matrix (Christensen et al., 2010), there are potential improvements if our method is applied to these estimators. Indeed, simulation and empirical results in Chapter 4.5 do suggest that pre-averaging can improve nonlinear shrinkage performance further. Comparisons with the thresholded version of these estimators (Kim et al., 2016) will then be revealing, and we leave these works in a future project.

p = 26 Methods	Out-of-Sample SD (%)	Aver Max Abs Weight(%)	Max Max Abs Weight(%)	Portfolio Turnover	Portfolio Return(%)	Sharpe Ratio
daily rebalancing portfolio with 5-day training window						
NERIVE	4.5	$20_{(6)}$	44	$0.26_{(0.08)}$	17.1	3.8
quarNERIVE	4.4	$19_{(5)}$	34	$0.23_{(0.07)}$	19.2	4.4
EQUAL	5.2	$4_{(-)}$	4	–	24.3	4.6
TSCV	5.9	$41_{(13)}$	92	$1.09_{(0.40)}$	16.5	2.8
GEC1	5.0	$30_{(11)}$	66	$0.33_{(0.14)}$	29.9	6.0
GEC2	5.0	$35_{(10)}$	80	$0.68_{(0.18)}$	19.0	3.8
GEC3	5.5	$39_{(11)}$	88	$0.95_{(0.26)}$	17.6	3.2
NORM0.1	4.6	$13_{(2)}$	18	$0.26_{(0.10)}$	17.6	3.8
NORM0.5	5.3	$33_{(6)}$	52	$0.83_{(0.25)}$	15.8	3.0
NORM1	5.7	$40_{(11)}$	76	$1.04_{(0.36)}$	15.8	2.7
TARVM	6.7	$51_{(37)}$	550	$1.53_{(1.59)}$	17.8	2.7
TS-POET	5.1	$30_{(8)}$	58	$0.60_{(0.2)}$	16.9	3.3
PR-NERIVE	4.4	$19_{(5)}$	34	$0.23_{(0.1)}$	20.1	4.6
PR-quarNERIVE	4.4	$20_{(5)}$	35	$0.22_{(0.1)}$	20.9	4.8
PR-POET	4.3	$25_{(6)}$	50	$0.33_{(0.1)}$	17.0	3.9
RPR-POET	7.9	$50_{(14)}$	110	$1.14_{(0.4)}$	17.5	2.2

Table 4.10 Empirical results (jumps removed) for the 26 most actively traded stocks in NYSE: 5-day training window

* Annualized out-of-sample standard deviation, averaged maximum absolute weight, maximum of maximum absolute weight, annualized portfolio return and Sharpe ratio for various methods, including GEC ($c = 1, 2, 3$) and NORM ($\delta = 0.1, 0.5, 1$).

p = 26 Methods	Out-of-Sample SD (%)	Aver Max Abs Weight(%)	Max Max Abs Weight(%)	Portfolio Turnover	Portfolio Return(%)	Sharpe Ratio
daily rebalancing portfolio with 1-day training window						
quarNERIVE	4.5	$18_{(7)}$	41	$0.58_{(0.19)}$	17.0	3.7
EQUAL	5.2	$4_{(-)}$	4	–	24.1	4.6
TSCV	6.5	$55_{(49)}$	708	$2.54_{(1.84)}$	23.4	3.6
GEC1	4.9	$28_{(12)}$	62	$0.55_{(0.16)}$	29.2	6.0
GEC2	5.0	$33_{(13)}$	84	$0.95_{(0.20)}$	28.5	5.8
GEC3	5.4	$36_{(13)}$	85	$1.32_{(0.31)}$	26.9	5.0
NORM0.1	4.6	$12_{(2)}$	20	$0.52_{(0.09)}$	20.9	4.5
NORM0.5	5.2	$30_{(6)}$	62	$1.34_{(0.23)}$	16.6	3.2
NORM1	5.7	$41_{(10)}$	73	$1.86_{(0.37)}$	17.8	3.1
TARVM	14.2	$42_{(90)}$	724	$0.38_{(20.56)}$	3.3	0.2
TS-POET	4.9	$28_{(11)}$	77	$1.25_{(0.4)}$	21.2	4.3
PR-quarNERIVE	4.5	$18_{(6)}$	40	$0.56_{(0.2)}$	17.3	3.8
PR-POET	4.7	$26_{(9)}$	63	$1.01_{(0.3)}$	23.7	5.0
RPR-POET	9.2	$65_{(27)}$	214	$2.51_{(0.7)}$	−9.2	−1.0

Table 4.11 Empirical results (jumps removed) for the 26 most actively traded stocks in NYSE: 1-day training window

* Annualized out-of-sample standard deviation, averaged maximum absolute weight, maximum of maximum absolute weight, annualized portfolio return and Sharpe ratio for various methods, including GEC ($c = 1, 2, 3$) and NORM ($\delta = 0.1, 0.5, 1$).

p = 73 Methods	Out-of-Sample SD (%)	Aver Max Abs Weight(%)	Max Max Abs Weight(%)	Portfolio Turnover	Portfolio Return(%)	Sharpe Ratio
daily rebalancing portfolio with 5-day training window						
NERIVE	3.9	$12_{(3)}$	22	$0.44_{(0.10)}$	14.9	3.9
quarNERIVE	3.9	$12_{(3)}$	21	$0.39_{(0.09)}$	15.3	3.9
EQUAL	5.4	$1_{(-)}$	1	$-$	22.3	4.1
TSCV	126.6	$236_{(503)}$	4693	$25.31_{(96.66)}$	-456.5	-3.6
GEC1	5.0	$20_{(10)}$	53	$0.34_{(0.14)}$	17.6	3.5
GEC2	4.9	$24_{(9)}$	60	$0.57_{(0.17)}$	13.2	2.7
GEC3	4.7	$25_{(9)}$	61	$0.87_{(0.23)}$	8.9	1.9
NORM0.1	4.4	$9_{(1)}$	16	$0.46_{(0.14)}$	15.0	3.4
NORM0.5	5.0	$20_{(5)}$	41	$1.33_{(0.34)}$	11.6	2.3
NORM1	6.1	$28_{(8)}$	61	$2.14_{(0.67)}$	8.8	1.4
TARVM	4.6	$5_{(2)}$	15	$0.09_{(0.03)}$	21.9	4.8
TS-POET	4.2	$19_{(5)}$	37	$0.87_{(0.3)}$	17.2	4.1
PR-NERIVE	3.8	$11_{(3)}$	22	$0.43_{(0.1)}$	15.9	4.1
PR-quarNERIVE	3.9	$11_{(3)}$	23	$0.40_{(0.1)}$	16.1	4.2
PR-POET	4.0	$16_{(5)}$	33	$0.56_{(0.2)}$	17.0	4.3
RPR-POET	5.9	$15_{(4)}$	31	$0.64_{(0.2)}$	23.4	4.0

Table 4.12 Empirical results (jumps removed) for the 73 most actively traded stocks in NYSE: 5-day training window

* Annualized out-of-sample standard deviation, averaged maximum absolute weight, maximum of maximum absolute weight, annualized portfolio return and Sharpe ratio for various methods, including GEC ($c = 1, 2, 3$) and NORM ($\delta = 0.1, 0.5, 1$).

p = 73 Methods	Out-of-Sample SD (%)	Aver Max Abs Weight(%)	Max Max Abs Weight(%)	Portfolio Turnover	Portfolio Return(%)	Sharpe Ratio
daily rebalancing portfolio with 1-day training window						
quarNERIVE	4.4	$8_{(3)}$	18	$0.77_{(0.19)}$	16.3	3.7
EQUAL	5.4	$1_{(-)}$	1	—	22.4	4.2
TSCV	85.5	$364_{(837)}$	7965	$42.63_{(846.86)}$	268.1	3.1
GEC1	5.2	$8_{(12)}$	69	$0.18_{(0.21)}$	20.4	3.9
GEC2	5.8	$20_{(13)}$	74	$0.86_{(0.21)}$	26.9	4.7
GEC3	10.2	$33_{(28)}$	157	$1.36_{(3.64)}$	3.0	0.3
NORM0.1	4.7	$8_{(2)}$	17	$0.78_{(0.16)}$	21.9	4.7
NORM0.5	8.6	$20_{(10)}$	51	$2.11_{(0.85)}$	29.2	3.4
NORM1	11.5	$28_{(16)}$	74	$3.20_{(2.31)}$	30.9	2.7
TARVM	332.8	$486_{(2383)}$	29145	$26.54_{(237.98)}$	684.5	2.1
TS-POET	5.4	$23_{(10)}$	82	$1.79_{(0.5)}$	20.7	3.9
PR-quarNERIVE	4.3	$8_{(2)}$	15	$0.81_{(0.2)}$	18.4	4.3
PR-POET	4.2	$13_{(5)}$	44	$0.96_{(0.3)}$	20.1	4.8
RPR-POET	6.5	$17_{(6)}$	49	$1.11_{(0.2)}$	21.0	3.2

Table 4.13 Empirical results (jumps removed) for the 73 most actively traded stocks in NYSE: 1-day training window

* Annualized out-of-sample standard deviation, averaged maximum absolute weight, maximum of maximum absolute weight, annualized portfolio return and Sharpe ratio for various methods, including GEC ($c = 1, 2, 3$) and NORM ($\delta = 0.1, 0.5, 1$).

4.7 Proof of Theorems

Before presenting the proofs, we present the last set of assumptions which are required for Theorem 4.1 to hold. We first need to decompose $\mathbf{X}_{v_s} - \mathbf{X}(s)$. Consider the previous-tick time $t_s^i \in (v_{s-1}, v_s]$ for the ith asset, which should satisfy

$$v_{s-1} < t_s^{(i_1)} \leq t_s^{(i_2)} \leq \cdots \leq t_s^{(i_p)} = v_s,$$

where $\{i_1, \ldots, i_p\}$ is some permutation of $1, \ldots, p$. Letting b_s denote the number of tides, we can write the above as

$$v_{s-1} < t_s^{j_1} < t_s^{j_2} < \cdots < t_s^{j_{p-b_s}} = v_s,$$

where $j_1, \ldots, j_{p-b_s} \in \{1, \ldots, p\}$.

Then we can write, for $s = 1, \ldots, nL$,

$$\mathbf{X}_{v_s} - \mathbf{X}(s) = \sum_{m=1}^{p-b_s-1} \mathbf{D}_m^s \mathbf{A}(t_s^{j_m}, t_s^{j_{m+1}}) \mathbf{Z}_{d,s}^j(m+1) + \sum_{m=1}^{p-b_s-1} \mathbf{D}_m^s \boldsymbol{\Sigma}(t_s^{j_m}, t_s^{j_{m+1}})^{1/2} \mathbf{Z}_{v,s}^j(m+1), \tag{4.15}$$

where $\mathbf{D}_m^s$ is a diagonal matrix with either 0 or 1 as elements. The jth diagonal element is 1 if the jth asset is already traded at time $t_s^{j_m}$, and 0 otherwise. The matrices $\mathbf{A}(\cdot,\cdot)$ and $\boldsymbol{\Sigma}(\cdot,\cdot)$ are as defined in Assumption (D1) and (V1) respectively.

(O3) The components of $\mathbf{Z}_{d,s}^j(m+1)$, $\mathbf{Z}_{v,s}^j(m+1) \in \mathcal{F}_{t_s^{j_{m+1}}}^j$ are conditionally independent given $\mathcal{F}_{-j}$, $E(\mathbf{Z}_{d,s}^j(m+1)|\mathcal{F}_{-j}) = \mathbf{0} = E(\mathbf{Z}_{v,s}^j(m+1)|\mathcal{F}_{-j})$, $\mathrm{var}(\mathbf{Z}_{d,s}^j(m+1)|\mathcal{F}_{-j}) = \mathbf{I}_p = \mathrm{var}(\mathbf{Z}_{v,s}^j(m+1)|\mathcal{F}_{-j})$ almost surely. Eighth moments exist for their components as well.
If the drift $\boldsymbol{\mu}_t$ is non-random, then $\mathbf{Z}_{d,s}^j(m+1) = (1, 0, \ldots 0)^{\mathrm{T}}$.

(O4) (Only for random drift.) Using notations in Assumption (D2), we assume that for some $c_{d,j,s} \in \mathcal{F}_{-j} \cup \mathcal{F}_s^j$ greater than 0, and for $\ell = 1, \ldots, m$,

$$\begin{aligned} &E\left(\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{D}_m^s \mathbf{A}(t_s^{j_m}, t_s^{j_{m+1}}) \mathbf{Z}_{d,s}^j(\ell+1) \middle| \mathcal{F}_{-j} \cup \mathcal{F}_{t_s^{j\ell}}^j\right) \\ &= \left(1 - \frac{c_{d,j,s}}{(p-b_s-1)^{1/6}}\right) \mathbf{p}_{ij}^{\mathrm{T}} \mathbf{D}_m^s \mathbf{A}(t_s^{j_m}, t_s^{j_{m+1}}) \mathbf{Z}_{d,s}^j(\ell) + e_{d,s}^{ij}(\ell), \end{aligned}$$

where we define $\mathbf{Z}^j_{d,s}(\ell)\mathbf{Z}^j_{d,s}(\ell)^{\mathrm{T}} = \mathbf{I}_p$ and $e^{ij}_{d,s}(\ell) = 0$ for $\ell \leq 0$. The process $\{e^{ij}_{d,s}(\ell)\}$ with $e^{ij}_{d,s}(\ell) \in \mathcal{F}^j_{t^{j_\ell}_s}$ has $E(e^{ij}_{d,s}(\ell)|\mathcal{F}_{-j} \cup \mathcal{F}^j_{t^{j_{\ell-1}}_s}) = 0$ almost surely, and $e^{ij}_{d,s}(\ell)|\mathcal{F}_{-j} \cup \mathcal{F}^j_{t^{j_{\ell-1}}_s} = O_P(\|\mathbf{A}(t^{j_{\ell-1}}_s, t^{j_\ell}_s)\|) = O_P(p^{1/2} \cdot (p - b_s - 1)^{-1} n^{-1} L^{-1})$.

The assumption for $E\Big(\mathbf{p}^{\mathrm{T}}_{ij}\mathbf{D}^s_m\mathbf{\Sigma}(t^{j_m}_s, t^{j_{m+1}}_s)^{1/2}\mathbf{Z}^j_{v,s}(\ell+1)\Big|\mathcal{F}_{-j} \cup \mathcal{F}^j_{t^{j_\ell}_s} \cup \mathcal{F}^{\boldsymbol{\sigma}}_{v_s}\Big)$ runs parallel to the above, with $c_{v,j,s} \in \mathcal{F}_{-j} \cup \mathcal{F}^j_s$ replaces $c_{d,j,s}$, $\mathbf{Z}^j_{v,s}(\cdot)$ replaces $\mathbf{Z}^j_{d,s}(\cdot)$, and $e^{ij}_{v,s}(\cdot)$ replaces $e^{ij}_{d,s}(\cdot)$ with

$$e^{ij}_{v,s}(\ell)|\mathcal{F}_{-j} \cup \mathcal{F}^j_{t^{j_{\ell-1}}_s} = O_P(\|\mathbf{\Sigma}(t^{j_{\ell-1}}_s, t^{j_\ell}_s)^{1/2}\|) = O_P(\|\mathbf{A}(t^{j_{\ell-1}}_s, t^{j_\ell}_s)\|/|t^{j_\ell}_s - t^{j_{\ell-1}}_s|^{1/2}).$$

(O5) (Only for random drift.) Let $\psi(x) = e^{x^2} - 1$. We assume that for $\ell = 1, \ldots, m$,

$$E\left\{\psi\left(\frac{|\mathbf{p}^{\mathrm{T}}_{ij}\mathbf{D}^s_m\mathbf{A}(t^{j_m}_s, t^{j_{m+1}}_s)\mathbf{Z}^j_{d,s}(\ell)|}{|\mathbf{p}^{\mathrm{T}}_{ij}\mathbf{D}^s_m\mathbf{A}(t^{j_m}_s, t^{j_{m+1}}_s)\mathbf{Z}^j_{d,s}(\ell-1)|}\right)\Big|\mathcal{F}_{-j} \cup \mathcal{F}^j_{t^{j_{\ell-1}}_s}\right\} < \infty,$$
$$E\left\{\psi\left(\frac{|e^{ij}_{d,s}(\ell)|}{|\mathbf{p}^{\mathrm{T}}_{ij}\mathbf{D}^s_m\mathbf{A}(t^{j_m}_s, t^{j_{m+1}}_s)\mathbf{Z}^j_{d,s}(\ell-1)|}\right)\Big|\mathcal{F}_{-j} \cup \mathcal{F}^j_{t^{j_{\ell-1}}_s}\right\} < \infty.$$

The assumption for the volatility runs parallel to the above, with the expectations now conditional on $\mathcal{F}_{-j} \cup \mathcal{F}^j_{t^{j_{\ell-1}}_s} \cup \mathcal{F}^{\boldsymbol{\sigma}}_{v_s}$ replaces $\mathbf{A}(\cdot,\cdot)$, $\mathbf{\Sigma}(\cdot,\cdot)^{1/2}$ replaces $\mathbf{A}(\cdot,\cdot)$, $\mathbf{Z}^j_{v,s}(\cdot)$ replaces $\mathbf{Z}^j_{d,s}(\cdot)$ and $e^{ij}_{v,s}(\cdot)$ replaces $e^{ij}_{d,s}(\cdot)$.

Assumptions (O3), (O4) and (O5) are parallel to (D1), (D2) and (D3) respectively. The major difference is that the coefficients $\rho^j_{d,K,q}, \rho^j_{v,K,q} \leq \xi < 1$ are now replaced by coefficients that are going to 1 as $n, p \to \infty$. This reflects that the correlations among variables between tick-by-tick trading times are high, since the time length between ticks is usually very small. Note that if the drift is non-random, we only need Assumption (O3) that $\mathbf{Z}^j_{d,s}(m+1) = (1, 0, \ldots, 0)^{\mathrm{T}}$, which is just a matter of notation rather than assumption.

We provide the proof of all the theorems of the chapter here. We assume the jump-diffusion model (4.10) for the log-price process $\{\mathbf{X}_t\}$, and prove Theorem 4.2, so that Theorem 4.1 then follows automatically. Define

$$\widetilde{\mathbf{Y}}_t = \mathbf{Y}_t - \widehat{\mathbf{J}}_t = (\mathbf{X}_t - \widehat{\mathbf{J}}_t) + \boldsymbol{\epsilon}_t = \widetilde{\mathbf{X}}_t + \boldsymbol{\epsilon}_t, \tag{4.16}$$

where $\{\widehat{\mathbf{J}}_t\}$ is the estimated jump process using the wavelet method in Fan and Wang (2007) described in Chapter 4.3.1. Then $\{\widetilde{\mathbf{X}}_t\}$ represents the jumps-removed log-price

process. For $j = 1, \ldots, L$ and $v_s = v_s^j$ for $s = 0, \ldots, n(j)$, we then have

$$\widetilde{\mathbf{Y}}(s) = \widetilde{\mathbf{X}}(s) + \boldsymbol{\epsilon}(s) = \widetilde{\mathbf{X}}_{v_s} + \mathbf{E}(s),$$

where we define

$$\mathbf{E}(s) = \boldsymbol{\epsilon}(s) + \widetilde{\mathbf{X}}(s) - \widetilde{\mathbf{X}}_{v_s} = \boldsymbol{\epsilon}(s) + (\mathbf{X}(s) - \widehat{\mathbf{J}}(s)) - (\mathbf{X}_{v_s} - \widehat{\mathbf{J}}_{v_s}).$$

We can then decompose, for $i = 1, \ldots, p$, $j = 1, \ldots, L$ with $\mathbf{P}_{-j} = (\mathbf{p}_{1j}, \ldots, \mathbf{p}_{pj})$,

$$\begin{aligned}\mathbf{p}_{ij}^{\mathrm{T}} \widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij} &= \mathbf{p}_{ij}^{\mathrm{T}} [\widetilde{\mathbf{Y}}, \widetilde{\mathbf{Y}}^{\mathrm{T}}]_j^{(K)} \mathbf{p}_{ij} - \frac{|S^j(K)|_K}{|S^j(1)|} \mathbf{p}_{ij}^{\mathrm{T}} [\widetilde{\mathbf{Y}}, \widetilde{\mathbf{Y}}^{\mathrm{T}}]_j^{(1)} \mathbf{p}_{ij} \\ &= I_1 + 2I_2 + I_3,\end{aligned}$$

where $\widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)$ is the TSCV in (4.4) constructed using jumps-removed data, and

$$\begin{aligned} I_1 &= \mathbf{p}_{ij}^{\mathrm{T}} [\widetilde{\mathbf{X}}_v, \widetilde{\mathbf{X}}_v^{\mathrm{T}}]_j^{(K)} \mathbf{p}_{ij} - \frac{|S^j(K)|_K}{|S^j(1)|} \mathbf{p}_{ij}^{\mathrm{T}} [\widetilde{\mathbf{X}}_v, \widetilde{\mathbf{X}}_v^{\mathrm{T}}]_j^{(1)} \mathbf{p}_{ij}, \\ I_2 &= \mathbf{p}_{ij}^{\mathrm{T}} [\widetilde{\mathbf{X}}_v, \mathbf{E}^{\mathrm{T}}]_j^{(K)} \mathbf{p}_{ij} - \frac{|S^j(K)|_K}{|S^j(1)|} \mathbf{p}_{ij}^{\mathrm{T}} [\widetilde{\mathbf{X}}_v, \mathbf{E}^{\mathrm{T}}]_j^{(1)} \mathbf{p}_{ij}, \\ I_3 &= \mathbf{p}_{ij}^{\mathrm{T}} [\mathbf{E}, \mathbf{E}^{\mathrm{T}}]_j^{(K)} \mathbf{p}_{ij} - \frac{|S^j(K)|_K}{|S^j(1)|} \mathbf{p}_{ij}^{\mathrm{T}} [\mathbf{E}, \mathbf{E}^{\mathrm{T}}]_j^{(1)} \mathbf{p}_{ij}, \end{aligned} \tag{4.17}$$

with $[\widetilde{\mathbf{X}}_v, \widetilde{\mathbf{X}}_v^{\mathrm{T}}]_j^{(m)}$, $[\widetilde{\mathbf{X}}_v, \mathbf{E}^{\mathrm{T}}]_j^{(m)}$ and $[\mathbf{E}, \mathbf{E}^{\mathrm{T}}]_j^{(m)}$ defined by

$$\begin{aligned}[\widetilde{\mathbf{X}}_v, \widetilde{\mathbf{X}}_v^{\mathrm{T}}]_j^{(m)} &= \frac{1}{m} \sum_{s, s+m \in S^j(m)} (\widetilde{\mathbf{X}}_{v_{s+m}} - \widetilde{\mathbf{X}}_{v_s})(\widetilde{\mathbf{X}}_{v_{s+m}} - \widetilde{\mathbf{X}}_{v_s})^{\mathrm{T}}, \\ [\widetilde{\mathbf{X}}_v, \mathbf{E}^{\mathrm{T}}]_j^{(m)} &= \frac{1}{m} \sum_{s, s+m \in S^j(m)} (\widetilde{\mathbf{X}}_{v_{s+m}} - \widetilde{\mathbf{X}}_{v_s})(\mathbf{E}(s+m) - \mathbf{E}(s))^{\mathrm{T}}, \\ [\mathbf{E}, \mathbf{E}^{\mathrm{T}}]_j^{(m)} &= \frac{1}{m} \sum_{s, s+m \in S^j(m)} (\mathbf{E}(s+m) - \mathbf{E}(s))(\mathbf{E}(s+m) - \mathbf{E}(s))^{\mathrm{T}}.\end{aligned}$$

Lemma 4.1 *Let all the assumptions in Theorem 4.2 hold. Then with $p/n \to c > 0$ when there are no pervasive factors, or $p^{3/2}/n \to c > 0$ when there are pervasive factors,*

$$\max_{\substack{i=1,\ldots,p \\ j=1,\ldots,L}} \left| \frac{I_1}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} - 1 \right| = O_P(n^{-1/6}).$$

Proof of Lemma 4.1. By Assumption (D1) and (V1), we first decompose for an integer $m \geq 1$, and $i = 1, \ldots, p$, $j = 1, \ldots, L$,

$$\begin{aligned}
&\mathbf{p}_{ij}^{\mathrm{T}}[\widetilde{\mathbf{X}}_v, \widetilde{\mathbf{X}}_v^{\mathrm{T}}]_j^{(m)}\mathbf{p}_{ij} = I_{11} + 2I_{12} + I_{13}, \quad \text{where} \\
I_{11} &= \frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=1}^{|S^j(m)|_m}(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{(r-1)m+q}, v_{rm+q})\mathbf{Z}_{d,rm+q}^j + \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(v_{(r-1)m+q}, v_{rm+q})^{1/2}\mathbf{Z}_{v,rm+q}^j)^2, \\
I_{12} &= \frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=1}^{|S^j(m)|_m}(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{(r-1)m+q}, v_{rm+q})\mathbf{Z}_{d,rm+q}^j + \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(v_{(r-1)m+q}, v_{rm+q})^{1/2}\mathbf{Z}_{v,rm+q}^j) \\
&\qquad \cdot (\mathbf{J}_{v_{rm+q}} - \widehat{\mathbf{J}}_{v_{rm+q}} - \mathbf{J}_{v_{(r-1)m+q}} + \widehat{\mathbf{J}}_{v_{(r-1)m+q}})^{\mathrm{T}}\mathbf{p}_{ij}, \\
I_{13} &= \frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=1}^{|S^j(m)|_m}\left((\mathbf{J}_{v_{rm+q}} - \widehat{\mathbf{J}}_{v_{rm+q}} - \mathbf{J}_{v_{(r-1)m+q}} + \widehat{\mathbf{J}}_{v_{(r-1)m+q}})^{\mathrm{T}}\mathbf{p}_{ij}\right)^2.
\end{aligned} \tag{4.18}$$

Consider further decomposition

$$\begin{aligned}
\left|\frac{I_{11}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}} - 1\right| &\leq \left|\frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=1}^{|S^j(m)|_m}(a_{d,r,m,q}^{ij}(r))^2\right| + \left|\frac{2}{m}\sum_{q=0}^{m-1}\sum_{r=1}^{|S^j(m)|_m}a_{d,r,m,q}^{ij}(r)b_{v,r,m,q}^{ij}(r)\right| \\
&\quad + \left|\frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=1}^{|S^j(m)|_m}(b_{v,r,m,q}^{ij}(r))^2 - 1\right|, \text{ where} \\
(a_{d,r,m,q}^{ij}(\ell))^2 &= (\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{(r-1)m+q}, v_{rm+q})\mathbf{Z}_{d,\ell m+q}^j)^2/\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}, \\
(b_{v,r,m,q}^{ij}(\ell))^2 &= (\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(v_{(r-1)m+q}, v_{rm+q})^{1/2}\mathbf{Z}_{v,\ell m+q}^j)^2/\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}.
\end{aligned}$$

To find the order of $I_{11}/\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij} - 1$, define

$$\begin{aligned}
g_{d,r,m,q}^{ij}(\ell) &= (a_{d,r,m,q}^{ij}(\ell))^2 - E((a_{d,r,m,q}^{ij}(\ell))^2|\mathcal{F}_{-j} \cup \mathcal{F}_{(\ell-1)m+q}^j), \\
g_{v,r,m,q}^{ij}(\ell) &= (b_{v,r,m,q}^{ij}(\ell))^2 - E((b_{v,r,m,q}^{ij}(\ell))^2|\mathcal{F}_{-j} \cup \mathcal{F}_{(\ell-1)m+q}^j \cup \mathcal{F}_{v_{rm+q}}^{\boldsymbol{\sigma}}).
\end{aligned}$$

Then we first consider

$$
\begin{aligned}
&\frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=1}^{|S^j(m)|_m}(a^{ij}_{d,r,m,q}(r))^2\\
&=\frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=1}^{|S^j(m)|_m}[(a^{ij}_{d,r,m,q}(r))^2-E((a^{ij}_{d,r,m,q}(r))^2|\mathcal{F}_{-j}\cup\mathcal{F}^j_{(r-1)m+q})]\\
&+\frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=1}^{|S^j(m)|_m}\left\{\rho^j_{d,m,q}(a^{ij}_{d,r,m,q}(r-1))^2\right.\\
&\left.+(1-\rho^j_{d,m,q})\frac{(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{(r-1)m+q},v_{rm+q}))^2}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}+\frac{e^{ij}_{d,(r-1)m+q}}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}\right\}\\
&=\frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=1}^{|S^j(m)|_m}g^{ij}_{d,r,m,q}(r)+\rho^j_{d,m,q}\cdot\frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=2}^{|S^j(m)|_m}g^{ij}_{d,r,m,q}(r-1)\\
&+\frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=1}^{|S^j(m)|_m}\left\{\frac{e^{ij}_{d,(r-1)m+q}}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}+\rho^j_{d,m,q}\cdot\frac{e^{ij}_{d,(r-2)m+q}}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}\right\}\\
&+(\rho^j_{d,m,q})^2\cdot\frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=3}^{|S^j(m)|_m}\left\{(a^{(ij)}_{d,r,m,q}(r-2))^2-\frac{(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{(r-1)m+q},v_{rm+q}))^2}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}\right\}\\
&+\frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=1}^{|S^j(m)|_m}\frac{(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{(r-1)m+q},v_{rm+q}))^2}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}\\
&=I_{11,1}+I_{11,2}+I_{11,3}+I_{11,4},
\end{aligned}
$$

where the equalities use Assumption (D2), and

$$
\begin{aligned}
I_{11,1}&=\frac{1}{m}\sum_{q=0}^{m-1}\left\{\sum_{\ell=0}^{\lfloor|S^j(m)|_m/2\rfloor-1}(\rho^j_{d,m,q})^\ell\sum_{r=1+\ell}^{|S^j(m)|_m}g^{ij}_{d,r,m,q}(r-\ell)\right\},\\
I_{11,2}&=\frac{1}{m}\sum_{q=0}^{m-1}\left\{\sum_{\ell=0}^{\lfloor|S^j(m)|_m/2\rfloor-1}(\rho^j_{d,m,q})^\ell\sum_{r=1+\ell}^{|S^j(m)|_m}\frac{e^{ij}_{d,(r-1-\ell)m+q}}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}\right\},\\
I_{11,3}&=(\rho^j_{d,m,q})^{\lfloor|S^j(m)|_m/2\rfloor}\\
&\quad\cdot\frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=\lfloor|S^j(m)|_m/2\rfloor+1}^{|S^j(m)|_m}\left\{(a^{(ij)}_{d,r,m,q}(r-2))^2-\frac{(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{(r-1)m+q},v_{rm+q}))^2}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}\right\},\\
I_{11,4}&=\frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=1}^{|S^j(m)|_m}\frac{(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{(r-1)m+q},v_{rm+q}))^2}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}.
\end{aligned}
$$

Letting $K^{ij}_{d,r,m,q}(\ell) = \frac{(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{(r-1)m+q}, v_{rm+q})\mathbf{Z}^j_{d,(\ell-1)m+q})^2}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}$, by Assumption (D2),

$$\begin{aligned} &E\left\{\psi\left(\frac{|g^{ij}_{d,r,m,q}(r-\ell)|}{K^{ij}_{d,r,m,q}(r-\ell)}\right)\Big|\mathcal{F}_{-j}\cup\mathcal{F}^j_{(r-1-\ell)m+q}\cup\mathcal{F}^{\boldsymbol{\sigma}}_{\tau_j}\right\}<\infty,\\ &E\left\{\psi\left(\frac{|e^{ij}_{d,(r-1-\ell)m+q}/\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}|}{K^{ij}_{d,r,m,q}(r-1-\ell)}\right)\Big|\mathcal{F}_{-j}\cup\mathcal{F}^j_{(r-2-\ell)m+q}\cup\mathcal{F}^{\boldsymbol{\sigma}}_{\tau_j}\right\}<\infty. \end{aligned} \tag{4.19}$$

At the same time, by Assumption (D1) that eighth moments exist for the $\mathbf{Z}^j_{d,(r-1-\ell)m+q}$'s and are conditionally independent given $\mathcal{F}_{-j}$, we can use Lemma 2.7 of Bai and Silverstein (1998) to arrive at

$$\begin{aligned} E((K^{ij}_{d,r,m,q}(r-\ell))^4|\mathcal{F}_{-j}\cup\mathcal{F}^{\boldsymbol{\sigma}}_{\tau_j}) &= O(\|\mathbf{A}(v_{(r-1)m+q}, v_{rm+q})\|^8/(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^4)\\ &= O(p_f m\cdot\frac{1}{(nL)^2}/\frac{1}{L})^4 = O(\frac{p_f m}{n^2 L})^2, \text{ so that}\\ K^{ij}_{d,r,m,q}(r-\ell)^2 &= O_P\Big(p_f m\cdot\frac{1}{(nL)^2}/\frac{1}{L}\Big)^2 = O_P\Big(\frac{p_f m}{n^2 L}\Big)^2, \end{aligned} \tag{4.20}$$

where the last line used Assumption (D1), with $p_f = 1$ if there are no pervasive factors and $p_f = p$ if there are pervasive factors or the drift is non-random, and the second line used Assumption (V1) on the rate of $\lambda_{\min}(\mathbf{\Sigma}(\tau_{j-1},\tau_j))$. With (4.19) and (4.20), we can apply Theorem 2.2 of van de Geer (2002) to arrive at

$$\sum_{r=1+\ell}^{|S^j(m)|_m} g^{ij}_{d,r,m,q}(r-\ell),\ \sum_{r=1+\ell}^{|S^j(m)|_m}\frac{e^{ij}_{d,(r-1-\ell)m+q}}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} = O_P\Big(|S^j(m)|_m^{1/2}\cdot\frac{p_f m}{n^2 L}\Big) = O_P\Big(\frac{p_f m^{1/2}}{n^{3/2}L}\Big),$$

for $\ell = 0, 1, \ldots, \lfloor |S^j(m)|_m/2\rfloor - 1$. Since $\rho^j_{d,m,q} \le \xi < 1$ uniformly by Assumption (D2), we have

$$I_{11,1},\ I_{11,2} = O_P\left(\frac{p_f m^{1/2}}{n^{3/2}L}\right). \tag{4.21}$$

Similar techniques in finding the order of $K^{ij}_{d,r,m,q}(r-\ell)$ show that

$$I_{11,3} = O_P\Big(\xi^{n/m}\cdot\frac{p_f m}{n^2 L}\Big). \tag{4.22}$$

For $I_{11,4}$, by (4.20), we have

$$I_{11,4} = O\Big(|S^j(m)|_m\cdot\frac{p_f m}{n^2 L}\Big) = O\Big(\frac{p_f}{nL}\Big). \tag{4.23}$$

Combining (4.21), (4.22) and (4.23), we have

$$\frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=1}^{|S^j(m)|_m}(a^{ij}_{d,r,m,q}(r))^2 = O_P(p_f n^{-1}L^{-1}). \tag{4.24}$$

Similar to the above calculations, by Assumption (V2), we can decompose

$$\begin{aligned}
&\frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=1}^{|S^j(m)|_m}(b^{ij}_{v,r,m,q}(r))^2 - 1 = J_{11,1} + J_{11,2} + J_{11,3} + J_{11,4}, \text{ where}\\
J_{11,1} &= \frac{1}{m}\sum_{q=0}^{m-1}\left\{\sum_{\ell=0}^{\lfloor |S^j(m)|_m/2\rfloor -1}(\rho^j_{v,m,q})^\ell \sum_{r=1+\ell}^{|S^j(m)|_m} g^{ij}_{v,r,m,q}(r-\ell)\right\},\\
J_{11,2} &= \frac{1}{m}\sum_{q=0}^{m-1}\left\{\sum_{\ell=0}^{\lfloor |S^j(m)|_m/2\rfloor -1}(\rho^j_{v,m,q})^\ell \sum_{r=1+\ell}^{|S^j(m)|_m} \frac{e^{ij}_{v,(r-1-\ell)m+q}}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}\right\},\\
J_{11,3} &= (\rho^j_{v,m,q})^{\lfloor |S^j(m)|_m/2\rfloor}\\
&\quad\cdot\frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=\lfloor |S^j(m)|_m/2\rfloor+1}^{|S^j(m)|_m}\left\{(b^{(ij)}_{v,r,m,q}(r-2))^2 - \frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(v_{(r-1)m+q},v_{rm+q})\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}\right\},\\
J_{11,4} &= \frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=1}^{|S^j(m)|_m}\frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(v_{(r-1)m+q},v_{rm+q})\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} - 1.
\end{aligned}$$

Letting $K^{ij}_{v,r,m,q}(\ell) = \frac{(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(v_{(r-1)m+q},v_{rm+q})^{1/2}\mathbf{Z}^j_{d,(\ell-1)m+q})^2}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}$, by Assumption (V2),

$$\begin{aligned}
&E\left\{\psi\left(\frac{|g^{ij}_{v,r,m,q}(r-\ell)|}{K^{ij}_{v,r,m,q}(r-\ell)}\right)\middle|\mathcal{F}_{-j}\cup\mathcal{F}^j_{(r-1-\ell)m+q}\right\} < \infty,\\
&E\left\{\psi\left(\frac{|e^{ij}_{v,(r-1-\ell)m+q}/\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}|}{K^{ij}_{v,r,m,q}(r-1-\ell)}\right)\middle|\mathcal{F}_{-j}\cup\mathcal{F}^j_{(r-2-\ell)m+q}\right\} < \infty.
\end{aligned} \tag{4.25}$$

At the same time, by Assumption (V1) that eighth moments exist for the $\mathbf{Z}^j_{v,(r-1-\ell)m+q}$'s and are conditionally independent given $\mathcal{F}_{-j}$, we can use Lemma 2.7 of Bai and Silverstein (1998) to arrive at

$$E((K^{ij}_{v,r,m,q}(r-\ell))^4|\mathcal{F}_{-j}\cup\mathcal{F}^{\sigma}_{\tau_j}) = O((\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(v_{(r-1)m+q},v_{rm+q})\mathbf{p}_{ij})^4/(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^4)$$
$$= O(p_f\cdot\frac{m}{nL}/\frac{p_f}{L})^4, \text{ so that}$$
$$K^{ij}_{v,r,m,q}(r-\ell)^2 = O_P\left(p_f\cdot\frac{m}{nL}/\frac{p_f}{L}\right)^2 = O_P\left(\frac{m}{n}\right)^2, \tag{4.26}$$

where the last line used Assumption (V1), with $p_f = 1$ if there are no pervasive factors and $p_f = p$ if there are pervasive factors. The main difference between (4.20) and (4.26) is that in (4.26), the numerator is a part of the denominator, and if pervasive factors affect the numerator, they have to affect the denominator too. This results in the balance of orders and hence p_f disappears from the order of the term. With (4.25) and (4.26), we can apply Theorem 2.2 of van de Geer (2002) to arrive at

$$J_{11,1},\ J_{11,2} = O_P\left(|S^j(m)|_m^{1/2}\cdot\frac{m}{n}\right) = O_P(m^{1/2}n^{-1/2}). \tag{4.27}$$

Similar to $I_{11,3}$, we have

$$J_{11,3} = O_P(\xi^{n/m}\cdot mn^{-1}). \tag{4.28}$$

For $J_{11,4}$, using Assumption (V1),

$$J_{11,4} = \frac{1}{m}\sum_{q=0}^{m-1}\frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(v_q,v_{n(j)-m+1+q})\mathbf{p}_{ij}-\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}$$
$$= -\frac{1}{m}\sum_{q=0}^{m-1}\frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(v_{n(j)-m+1+q},\tau_j)\mathbf{p}_{ij}+\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},v_q)\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}$$
$$= O_P\left(\frac{1}{m}\sum_{q=0}^{m-1}\frac{(m-1-q)+q}{nL}/\frac{1}{L}\right) = O_P(mn^{-1}). \tag{4.29}$$

Combining (4.27), (4.28) and (4.29), we have

$$\frac{1}{m}\sum_{q=0}^{m-1}\sum_{r=1}^{|S^j(m)|_m}(b^{ij}_{v,r,m,q}(r))^2 - 1 = O_P(m^{1/2}n^{-1/2}). \tag{4.30}$$

Using the Cauchy-Schwarz inequality, using (4.24) and (4.30), we have

$$\left|\frac{I_{11}}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}-1\right| = O_P(p_f n^{-1}L^{-1}+m^{1/2}n^{-1/2}+p_f^{1/2}n^{-1/2}L^{-1/2}) = O_P(n^{-1/6}), \tag{4.31}$$

if there are pervasive factors such that $p_f = p \asymp n^{2/3}$ and $m = O(n^{2/3})$. Turning to I_{12} and I_{13} defined in (4.17), using Assumption (W1) to (W3), and the rate in Fan and Wang (2007), we have

$$I_{13}/\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij} = O_P(n^{-1/2}L^{1/2}).$$

The above implies, through using the Cauchy-Schwarz inequality,

$$I_{12}/\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij} = O_P(n^{-1/4}L^{1/4}).$$

Combining all results, we have for $K \asymp n^{2/3}$,

$$\left|\frac{\mathbf{p}_{ij}^{\mathrm{T}}[\widetilde{\mathbf{X}}_v,\widetilde{\mathbf{X}}_v^{\mathrm{T}}]_j^{(K)}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}-1\right| = O_P(n^{-1/6}),$$
$$\frac{|S^j(K)|_K}{|S^j(1)|}\cdot\left|\frac{\mathbf{p}_{ij}^{\mathrm{T}}[\widetilde{\mathbf{X}}_v,\widetilde{\mathbf{X}}_v^{\mathrm{T}}]_j^{(1)}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}\right| = O_p(K^{-1}\cdot 1) = O_P(n^{-2/3}).$$

Note that the above bounds are independent of the indices i and j, and hence

$$\max_{\substack{i=1,\ldots,p\\ j=1,\ldots,L}}\left|\frac{I_1}{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}-1\right| = O_P(n^{-1/6}+n^{-2/3}) = O_P(n^{-1/6}).$$

This completes of proof of the lemma. □

Lemma 4.2 *Let all the assumptions in Theorem 4.2 hold. Then with $p/n \to c > 0$ when there are no pervasive factors, or $p^{3/2}/n \to c > 0$ when there are pervasive factors,*

$$\max_{\substack{i=1,\ldots,p\\ j=1,\ldots,L}}\max_{s=1,\ldots,n(j)}\left|\frac{\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}_{v_s}-\mathbf{X}(s))}{(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}}\right| = O_P(p^{1/6}n^{-1/2}).$$

Proof of Lemma 4.2. Consider $\frac{\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}_{v_s}-\mathbf{X}(s))}{(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}} = A_d^{ij}(s) + A_v^{ij}(s)$, where using (4.15),

$$A_d^{ij}(s) = \sum_{m=1}^{p-b_s-1} \frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{D}_m^s\mathbf{A}(t_s^{j_m}, t_s^{j_{m+1}})\mathbf{Z}_{d,s}^j(m+1)}{(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}},$$
$$A_v^{ij}(s) = \sum_{m=1}^{p-b_s-1} \frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{D}_m^s\boldsymbol{\Sigma}(t_s^{j_m}, t_s^{j_{m+1}})^{1/2}\mathbf{Z}_{v,s}^j(m+1)}{(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}}.$$

We first deal with non-random drift for $A_d^{ij}(s)$. By Assumption (D1) and (V1), we have

$$\begin{aligned}|A_d^{ij}(s)| &\le \sum_{m=1}^{p-b_s-1} \frac{\|\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{D}_m^s\mathbf{A}(t_s^{j_m}, t_s^{j_{m+1}})\|}{(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}} \\ &= O_P((p-b_s-1)\cdot p^{1/2}\cdot(p-b_s-1)^{-1}n^{-1}L^{-1}/L^{1/2}) \\ &= O_P(p^{1/2}n^{-1}). \end{aligned} \tag{4.32}$$

Now we focus on random drift. Define for $\ell = 1, \ldots, m+1$,

$$g_{d,s,m}^{ij}(\ell) = \frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{D}_m^s\mathbf{A}(t_s^{j_m}, t_s^{j_{m+1}})\mathbf{Z}_{d,s}^j(\ell) - E(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{D}_m^s\mathbf{A}(t_s^{j_m}, t_s^{j_{m+1}})\mathbf{Z}_{d,s}^j(\ell)|\mathcal{F}_{-j}\cup\mathcal{F}_{t_s^{j_{\ell-1}}}^j)}{(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}},$$
$$g_{v,s,m}^{ij}(\ell) = \frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{D}_m^s\boldsymbol{\Sigma}(t_s^{j_m}, t_s^{j_{m+1}})^{1/2}\mathbf{Z}_{v,s}^j(\ell) - E(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{D}_m^s\boldsymbol{\Sigma}(t_s^{j_m}, t_s^{j_{m+1}})^{1/2}\mathbf{Z}_{v,s}^j(\ell)|\mathcal{F}_{-j}\cup\mathcal{F}_{t_s^{j_{\ell-1}}}^j\cup\mathcal{F}_{\tau_j}^{\boldsymbol{\sigma}})}{(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}}.$$

Consider $A_d^{ij}(s)$ first. By Assumption (O4), we can decompose

$$\begin{aligned}A_d^{ij}(s) &= \sum_{m=1}^{p-b_s-1} g_{d,s,m}^{ij}(m+1) + \left(1 - \frac{c_{d,j,s}}{(p-b_s-1)^{1/6}}\right)\sum_{m=1}^{p-b_s-1}\frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{D}_m^s\mathbf{A}(t_s^{j_m}, t_s^{j_{m+1}})\mathbf{Z}_{d,s}^j(m)}{(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}} \\ &\quad + \sum_{m=1}^{p-b_s-1}\frac{e_{d,s}^{ij}(m)}{(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^{1/2}} \\ &= J_1 + J_2 + J_3,\end{aligned}$$

where

$$
\begin{aligned}
J_1 &= \sum_{\ell=0}^{\lfloor (p-b_s-1)/2 \rfloor - 1} \left(1 - \frac{c_{d,j,s}}{(p-b_s-1)^{1/6}}\right)^{\ell} \sum_{m=1}^{p-b_s-1} g_{d,s,m}^{ij}(m-\ell+1), \\
J_2 &= \sum_{\ell=0}^{\lfloor (p-b_s-1)/2 \rfloor - 1} \left(1 - \frac{c_{d,j,s}}{(p-b_s-1)^{1/6}}\right)^{\ell} \sum_{m=1}^{p-b_s-1} \frac{e_{d,s}^{ij}(m-\ell)}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij})^{1/2}}, \\
J_3 &= \left(1 - \frac{c_{d,j,s}}{(p-b_s-1)^{1/6}}\right)^{\lfloor (p-b_s-1)/2 \rfloor} \\
&\quad \cdot \sum_{m=1}^{p-b_s-1} \frac{\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{D}_m^s \mathbf{A}(t_s^{j_m}, t_s^{j_{m+1}}) \mathbf{Z}_{d,s}^j (m + 1 - \lfloor (p-b_s-1)/2 \rfloor)}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij})^{1/2}}.
\end{aligned}
$$

Letting $K_{d,s,m}^{ij}(\ell) = \frac{|\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{D}_m^s \mathbf{A}(t_s^{j_m}, t_s^{j_{m+1}}) \mathbf{Z}_{d,s}^j(\ell-1)|}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij})^{1/2}}$, by Assumption (O5),

$$
\begin{aligned}
&E\left\{\psi\left(\frac{|g_{d,s,m}^{ij}(m-\ell+1)|}{K_{d,s,m}^{ij}(m-\ell+1)}\right) \middle| \mathcal{F}_{-j} \cup \mathcal{F}_{t_s^{j_{m-\ell}}}^j \right\} < \infty, \\
&E\left\{\psi\left(\frac{|e_{d,(s}^{ij}(m-\ell)/(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij})^{1/2}|}{K_{d,s,m}^{ij}(m-\ell)}\right) \middle| \mathcal{F}_{-j} \cup \mathcal{F}_{t_s^{j_{m-\ell-1}}}^j \right\} < \infty.
\end{aligned} \tag{4.33}
$$

At the same time, by Assumption (O3) that fourth moments exist for the $\mathbf{Z}_{d,s}^j(\ell)$'s and are conditionally independent given $\mathcal{F}_j$, we can use Lemma 2.7 of Bai and Silverstein (1998) to arrive at

$$
\begin{aligned}
E(K_{d,s,m}^{ij}(m-\ell+1)^4 | \mathcal{F}_{-j}) &= O(\|\mathbf{A}(t_s^{j_m}, t_s^{j_{m+1}})\|^4 / (\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij})^2) \\
&= O(p_f \cdot (p-b_s-1)^{-2} n^{-2} L^{-2} / L^{-1}) \\
&= O(p_f \cdot (p-b_s-1)^{-2} n^{-2} L^{-1}), \text{ so that} \\
K_{d,s,m}^{ij}(m-\ell+1)^2 &= O_P(p_f \cdot (p-b_s-1)^{-2} n^{-2} L^{-1}),
\end{aligned} \tag{4.34}
$$

where $p_f = 1$ if there are no pervasive factors and $p_f = p$ if there are pervasive factors. With (4.33) and (4.34), we can apply Theorem 2.2 of van de Geer (2002) to arrive at

$$
\begin{aligned}
&\sum_{m=1}^{p-b_s-1} g_{d,s,m}^{ij}(m-\ell-1), \sum_{m=1}^{p-b_s-1} \frac{e_{d,s}^{ij}(m-\ell)}{(\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij})^{1/2}} \\
&= O_P(p^{1/2} \cdot p_f^{1/2} \cdot (p-b_s-1)^{-1} n^{-1} L^{-1/2}) \\
&= O_P(p_f^{1/2} p^{-1/2} n^{-1} L^{-1/2}).
\end{aligned}
$$

The above implies that

$$J_1, J_2 = O_P((p - b_s - 1)^{1/6} \cdot p_f^{1/2} p^{-1/2} n^{-1} L^{-1/2}) = O_P(p_f^{1/2} p^{-1/3} n^{-1} L^{-1/2}).$$

We also have, as $p \to \infty$,

$$J_3 = O_P(e^{-c_{d,j,s} p^{5/6}/2} p_f^{1/2} n^{-1} L^{-1/2}).$$

The above results give

$$A_d^{ij}(s) = O_P(p_f^{1/2} p^{-1/3} n^{-1} L^{-1/2}). \tag{4.35}$$

Parallel arguments show that

$$\begin{aligned}
&\sum_{m=1}^{p-b_s-1} g_{v,s,m}^{ij}(m - \ell - 1), \sum_{m=1}^{p-b_s-1} \frac{e_{v,s}^{ij}(m-\ell)}{(\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij})^{1/2}} \\
&= O_P((p - b_s - 1)^{1/2} \cdot (\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{D}_m^s \mathbf{\Sigma}(t_s^{j_m}, t_s^{j_{m+1}}) \mathbf{D}_m^s \mathbf{p}_{ij})^{1/2} / (\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij})^{1/2}) \\
&= O_P((p - b_s - 1)^{1/2} \cdot (p - b_s - 1)^{-1/2} n^{-1/2} L^{-1/2} / L^{-1/2}) \\
&= O_P(n^{-1/2}),
\end{aligned}$$

where p_f cancels since $\mathbf{D}_m^s$ is only a diagonal matrix of 1 or 0, and hence if pervasive factors are affecting the numerator, it has to affect the denominator too. Parallel arguments as before show that

$$A_v^{ij}(s) = O_P(p^{1/6} n^{-1/2}). \tag{4.36}$$

Combining (4.32), (4.35) and (4.36), since we at most have $p^{3/2}/n \to c > 0$,

$$\frac{\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}_{v_s} - \mathbf{X}(s))}{(\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij})^{1/2}} = O_P(p^{1/6} n^{-1/2} + p^{1/2} n^{-1}) = O_P(p^{1/6} n^{-1/2}). \tag{4.37}$$

This completes the proof of the theorem, since the above rate is free of all indices. □

Lemma 4.3 *Let all the assumptions in Theorem 4.2 hold. Then with $p/n \to c > 0$ when there are no pervasive factors, or $p^{3/2}/n \to c > 0$ when there are pervasive factors,*

$$\max_{\substack{i=1,\ldots,p \\ j=1,\ldots,L}} \left| \frac{I_2}{\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}} \right| = O_P(n^{-1/6}).$$

Proof of Lemma 4.3 For an integer $m \geq 1$, $i = 1, \ldots, p$ and $j = 1, \ldots, L$, write

$$\frac{\mathbf{p}_{ij}^{\mathrm{T}}[\widetilde{\mathbf{X}}_v, \mathbf{E}^{\mathrm{T}}]_j^{(m)}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}} = \sum_{i=1}^{3}(I_{2,i} + J_i + K_i), \text{ where, defining } e(\mathbf{J}_t) = \mathbf{J}_t - \widehat{\mathbf{J}}_t,$$

$$I_{2,1} = \frac{1}{m}\sum_{s,s-m\in S^j(m)} \frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-m}, v_s)\mathbf{Z}_{d,s}^j(\boldsymbol{\epsilon}(s) - \boldsymbol{\epsilon}(s-m))^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}},$$

$$I_{2,2} = \frac{1}{m}\sum_{s,s-m\in S^j(m)} \frac{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(v_{s-m}, v_s)^{1/2}\mathbf{Z}_{v,s}^j(\boldsymbol{\epsilon}(s) - \boldsymbol{\epsilon}(s-m))^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}},$$

$$I_{2,3} = \frac{1}{m}\sum_{s,s-m\in S^j(m)} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(e(\mathbf{J}_{v_s}) - e(\mathbf{J}_{v_{s-m}}))(\boldsymbol{\epsilon}(s) - \boldsymbol{\epsilon}(s-m))^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}},$$

$$J_1 = \frac{1}{m}\sum_{s,s-m\in S^j(m)} \frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-m}, v_s)\mathbf{Z}_{d,s}^j(\mathbf{X}(s) - \mathbf{X}_{v_s} + \mathbf{X}_{v_{s-m}} - \mathbf{X}(s-m))^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}},$$

$$J_2 = \frac{1}{m}\sum_{s,s-m\in S^j(m)} \frac{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(v_{s-m}, v_s)^{1/2}\mathbf{Z}_{v,s}^j(\mathbf{X}(s) - \mathbf{X}_{v_s} + \mathbf{X}_{v_{s-m}} - \mathbf{X}(s-m))^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}},$$

$$J_3 = \frac{1}{m}\sum_{s,s-m\in S^j(m)} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(e(\mathbf{J}_{v_s}) - e(\mathbf{J}_{v_{s-m}})(\mathbf{X}(s) - \mathbf{X}_{v_s} + \mathbf{X}_{v_{s-m}} - \mathbf{X}(s-m))^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}},$$

$$K_1 = \frac{1}{m}\sum_{s,s-m\in S^j(m)} \frac{\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-m}, v_s)\mathbf{Z}_{d,s}^j(e(\mathbf{J}(s)) - e(\mathbf{J}_{v_s}) - e(\mathbf{J}(s-m)) + e(\mathbf{J}_{v_{s-m}}))^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}},$$

$$K_2 = \frac{1}{m}\sum_{s,s-m\in S^j(m)} \frac{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(v_{s-m}, v_s)^{1/2}\mathbf{Z}_{v,s}^j(e(\mathbf{J}(s)) - e(\mathbf{J}_{v_s}) - e(\mathbf{J}(s-m)) + e(\mathbf{J}_{v_{s-m}}))^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}},$$

$$K_3 = \frac{1}{m}\sum_{s,s-m\in S^j(m)} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(e(\mathbf{J}_{v_s}) - e(\mathbf{J}_{v_{s-m}}))(e(\mathbf{J}(s)) - e(\mathbf{J}_{v_s}) - e(\mathbf{J}(s-m)) + e(\mathbf{J}_{v_{s-m}}))^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}}.$$

Consider $g_{d,s}^{ij} = \mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-m}, v_s)\mathbf{Z}_{d,s}^j\boldsymbol{\epsilon}(s)^{\mathrm{T}}\mathbf{p}_{ij}$. Then

$$E\left(\left(\frac{1}{m}\sum_{s,s-m\in S^j(m)} g_{d,s}^{ij}\right)^2 \middle| \mathcal{F}_{-j}\right) = \frac{1}{m^2}\sum_{s,s-m\in S^j(m)} E((g_{d,s}^{ij})^2|\mathcal{F}_{-j}) + \frac{1}{m^2}\sum_{\substack{s_k, s_k+m\in S^j(m)\\ s_1\neq s_2}} E(g_{d,s_1}^{ij}g_{d,s_2}^{ij}|\mathcal{F}_{-j}). \tag{4.38}$$

With Assumption (D1) and (E2), we can use Lemma 2.7 of Bai and Silverstein (1998) to arrive at

$$\begin{aligned}E((g_{d,s}^{ij})^2|\mathcal{F}_{-j}) &\leq E^{1/2}((\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-m},v_s)\mathbf{Z}_{d,s}^{j})^4|\mathcal{F}_{-j})E^{1/2}((\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{\Sigma}_{\epsilon,s}^{j})^{1/2}\mathbf{Z}_{\epsilon,s}^{j})^4|\mathcal{F}_{-j})\\ &= O(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s-m},v_s)\mathbf{A}(v_{s-m},v_s)^{\mathrm{T}}\mathbf{p}_{ij}\cdot E^{1/2}((\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}_{\epsilon,s}^{j}\mathbf{p}_{ij})^2|\mathcal{F}_{-j}))\\ &= O(\|\mathbf{A}(v_{s-m},v_s)\|^2\cdot\lambda_\epsilon) = O_P(p_f mn^{-2}L^{-2}),\end{aligned}$$

where $p_f = p$ if there are pervasive factors, and $p_f = 1$ otherwise. Also, by Assumption (E3), since $E(\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s)|\mathcal{F}_{-j}) = 0$, by Theorem 1.4 in Rio (2013) we have that

$$\begin{aligned}E(g_{d,s_1}^{ij}g_{v,s_2}^{ij}|\mathcal{F}_{-j}) &\leq 2O(n^{-1})E^{1/2}((\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s_1))^2|\mathcal{F}_{-j})\\ &\quad\cdot E^{1/2}((\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s_2)\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s_1-m},v_{s_1})\mathbf{Z}_{d,s_1}^{j}\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s_2-m},v_{s_2})\mathbf{Z}_{d,s_2}^{j})^2|\mathcal{F}_{-j})\\ &\leq 2O(n^{-1})E^{1/2}((\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s_1))^2|\mathcal{F}_{-j})\cdot E^{1/4}((\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s_2))^4|\mathcal{F}_{-j})\\ &\quad\cdot E^{1/8}((\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s_1-m},v_{s_1})\mathbf{Z}_{d,s_1}^{j})^8|\mathcal{F}_{-j})E^{1/8}((\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{A}(v_{s_2-m},v_{s_2})\mathbf{Z}_{d,s_2}^{j})^8|\mathcal{F}_{-j})\\ &= O(n^{-1}\|\mathbf{A}(v_{s-m},v_s)\|^2) = O(p_f mn^{-3}L^{-2}),\end{aligned}$$

where the third inequality sign used Lemma 2.7 of Bai and Silverstein (1998). Using these two results, (4.38) becomes

$$E\left(\left(\frac{1}{m}\sum_{s,s-m\in S^j(m)}g_{d,s}^{ij}\right)^2\middle|\mathcal{F}_{-j}\right) = O(m^{-2}p_f mn^{-1}L^{-2}) = O(p_f m^{-1}n^{-1}L^{-2}).$$

This implies that

$$I_{2,1} = O_P(p_f^{1/2}m^{-1/2}n^{-1/2}L^{-1}/L^{-1}) = O_P(p_f^{1/2}m^{-1/2}n^{-1/2}). \tag{4.39}$$

Now consider $g_{v,s}^{ij} = \mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(v_{s-m},v_s)^{1/2}\mathbf{Z}_{v,s}^{j}\boldsymbol{\epsilon}(s)^{\mathrm{T}}\mathbf{p}_{ij}/\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}$. Parallel arguments using Assumption (V1) and (E2) give

$$\begin{aligned}E((g_{v,s}^{ij})^2|\mathcal{F}_{-j}\cup\mathcal{F}_{\tau_j}^{\boldsymbol{\sigma}}) &= O(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(v_{s-m},v_s)\mathbf{p}_{ij}/(\mathbf{p}_{ij}^{\mathrm{T}}\mathbf{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij})^2) = O(p_f mn^{-1}L^{-1}/(p_f^2L^{-2}))\\ &= O(p_f^{-1}mn^{-1}L),\\ E(g_{v,s_1}^{ij}g_{v,s_2}^{ij}|\mathcal{F}_{-j}\cup\mathcal{F}_{\tau_j}^{\boldsymbol{\sigma}}) &= O(n^{-1}p_f^{-1}mn^{-1}L) = O(p_f^{-1}mn^{-2}L)).\end{aligned}$$

Hence using decomposition parallel to (4.38),

$$I_{2,2} = O_P(m^{-2}\cdot p_f^{-1}mL)^{1/2} = O_P(p_f^{-1/2}m^{-1/2}L^{1/2}). \tag{4.40}$$

For terms involving jumps, using Assumption (W1) to (W3) and the rate in Fan and Wang (2007), we have

$$\begin{aligned}
I_{2,3} &= O_P(n^{-1/4}L^{3/4}),\\
J_3 &= O_P(n^{-1/4}L^{1/4}\cdot p^{1/6}n^{-1/2}) = O_P(p^{1/6}n^{-3/4}L^{1/4}),\\
K_1 &= O_P(\|\mathbf{A}(v_{s-m},v_s)\|/L^{-1}\cdot n^{-1/4}L^{-1/4}) = O_P(p_f^{1/2}m^{1/2}n^{-5/4}L^{-1/4}),\\
K_2 &= O_P(p_f^{-1/2}m^{1/2}n^{-1/2}L^{1/2}\cdot n^{-1/4}L^{-1/4}) = O_P(p_f^{-1/2}m^{1/2}n^{-3/4}L^{1/4}),\\
K_3 &= O_P(n^{-1/2}L^{1/2}),
\end{aligned} \tag{4.41}$$

where J_3 used the result of Lemma 4.2. Using the result of Lemma 4.2 again, we have

$$\begin{aligned}
J_1 &= O_P(nm^{-1}\cdot p_f^{1/2}m^{1/2}n^{-1}L^{-1/2}\cdot p^{1/6}n^{-1/2}) = O_P(p_f^{1/2}p^{1/6}m^{-1/2}n^{-1/2}L^{-1/2}),\\
J_2 &= O_P(nm^{-1}\cdot m^{1/2}n^{-1/2}\cdot p^{1/6}n^{-1/2}) = O_P(m^{-1/2}p^{1/6}).
\end{aligned} \tag{4.42}$$

At $m = K \asymp n^{2/3}$, (4.39), (4.40), (4.41) and (4.42) imply that, for $p_f = 1$ with $p \asymp n$ or $p_f = p \asymp n^{2/3}$,

$$\frac{\mathbf{p}_{ij}^{\mathrm{T}}[\widetilde{\mathbf{X}}_v, \mathbf{E}^{\mathrm{T}}]_j^{(K)}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} = O_P(n^{-1/6}).$$

At $m = 1$, (4.39), (4.40), (4.41) and (4.42) imply that, for $p_f = 1$ with $p \asymp n$ or $p_f = p \asymp n^{2/3}$,

$$\frac{\mathbf{p}_{ij}^{\mathrm{T}}[\widetilde{\mathbf{X}}_v, \mathbf{E}^{\mathrm{T}}]_j^{(1)}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}} = O_P(p^{1/6}).$$

Since the above two results are free of all indices, they imply that

$$\max_{\substack{i=1,\ldots,p\\ j=1,\ldots,L}}\left|\frac{I_2}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}\right| = O_P(n^{-1/6}+p^{1/6}K^{-1}) = O_P(n^{-1/6}).$$

This completes the proof of the lemma. □

Lemma 4.4 *Let all the assumptions in Theorem 4.2 hold. Then with $p/n \to c > 0$ when there are no pervasive factors, or $p^{3/2}/n \to c > 0$ when there are pervasive factors,*

$$\max_{\substack{i=1,\ldots,p\\ j=1,\ldots,L}}\left|\frac{I_3}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1},\tau_j)\mathbf{p}_{ij}}\right| = O_P(n^{-1/6}).$$

Proof of Lemma 4.4. Consider for an integer $m \geq 1$ and $i = 1, \ldots, p$, $j = 1, \ldots, L$, using the notations in the proof of Lemma 4.3,

$$\frac{\mathbf{p}_{ij}^{\mathrm{T}}[\mathbf{E}, \mathbf{E}^{\mathrm{T}}]_j^{(m)}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}} = \sum_{\ell=1}^{3} I_{3,\ell} + 2\sum_{\ell=1}^{3} I_{3,\ell}, \text{ where}$$

$$I_{3,1}(m) = \frac{1}{m}\sum_{s, s-m \in S^j(m)} \frac{(\mathbf{p}_{ij}^{\mathrm{T}}(\boldsymbol{\epsilon}(s) - \boldsymbol{\epsilon}(s-m)))^2}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}},$$

$$I_{3,2} = \frac{1}{m}\sum_{s, s-m \in S^j(m)} \frac{(\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}(s) - \mathbf{X}_{v_s} + \mathbf{X}_{v_{s-m}} - \mathbf{X}(s-m)))^2}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}},$$

$$I_{3,3} = \frac{1}{m}\sum_{s, s-m \in S^j(m)} \frac{(\mathbf{p}_{ij}^{\mathrm{T}}(e(\mathbf{J}(s)) - e(\mathbf{J}_{v_s}) - e(\mathbf{J}(s-m)) + e(\mathbf{J}_{v_{s-m}})))^2}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}},$$

$$I_{3,4} = \frac{1}{m}\sum_{s, s-m \in S^j(m)} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(\boldsymbol{\epsilon}(s) - \boldsymbol{\epsilon}(s-m))(\mathbf{X}(s) - \mathbf{X}_{v_s} + \mathbf{X}_{v_{s-m}} - \mathbf{X}(s-m))^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}},$$

$$I_{3,5} = \frac{1}{m}\sum_{s, s-m \in S^j(m)} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(\boldsymbol{\epsilon}(s) - \boldsymbol{\epsilon}(s-m))(e(\mathbf{J}(s)) - e(\mathbf{J}_{v_s}) - e(\mathbf{J}(s-m)) + e(\mathbf{J}_{v_{s-m}}))^{\mathrm{T}}\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}},$$

$$I_{3,6} = \frac{1}{m}\sum_{s, s-m \in S^j(m)} \frac{\mathbf{p}_{ij}^{\mathrm{T}}(\mathbf{X}(s) - \mathbf{X}_{v_s} + \mathbf{X}_{v_{s-m}} - \mathbf{X}(s-m))}{\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}} \cdot (e(\mathbf{J}(s)) - e(\mathbf{J}_{v_s}) - e(\mathbf{J}(s-m)) + e(\mathbf{J}_{v_{s-m}}))^{\mathrm{T}}\mathbf{p}_{ij}.$$

We consider $I_{3,2}$ first, which by Lemma 4.2 has

$$I_{3,2} = O_P(nm^{-1} \cdot p^{1/3}n^{-1}) = O_P(p^{1/3}m^{-1}).$$

Using Assumption (W1) to (W3) and the rate of wavelet removal in Fan and Wang (2007), we have

$$I_{3,3} = O_P(n^{-1/2}L^{1/2}),$$
$$I_{3,5} = O_P(n^{-1/4}L^{3/4}),$$
$$I_{3,6} = O_P(p^{1/6}n^{-1/2} \cdot n^{-1/4}L^{1/4}) = O_P(p^{1/6}n^{-3/4}L^{1/4}).$$

Consider $h_s^{ij} = \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s)(\mathbf{X}(s) - \mathbf{X}_{v_s})^{\mathrm{T}}\mathbf{p}_{ij}/\mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}$. Then using Assumption (E3), (D1) and (V1) that eighth moments exist, with $s_1 \neq s_2$,

$$E((h_s^{ij})^2|\mathcal{F}_{-j} \cup \mathcal{F}_{\tau_j}^{\boldsymbol{\sigma}}) = O(p^{1/3}n^{-1}L),$$
$$E(h_{s_1}^{ij}h_{s_2}^{ij}|\mathcal{F}_{-j} \cup \mathcal{F}_{\tau_j}^{\boldsymbol{\sigma}}) = O(n^{-1} \cdot L \cdot p^{1/3}n^{-1}) = O(p^{1/3}n^{-2}L).$$

Hence using decomposition parallel to (4.38), we can conclude that

$$I_{3,4} = O_P(m^{-2} \cdot n \cdot p^{1/3} n^{-1} L + m^{-2} \cdot n^2 \cdot p^{1/3} n^{-2} L)^{1/2} = O_P(p^{1/6} m^{-1} L^{1/2}).$$

Finally, for $K \asymp n^{2/3}$, we consider the rate of

$$(\mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{p}_{ij}) \left(I_{3,1}(K) - \frac{|S^j(K)|_K}{|S^j(1)|} I_{3,1}(1) \right) = J_1 - 2J_2 + J_3, \text{ where}$$

$$J_1 = \frac{1}{K} \sum_{s, s-K \in S^j(K)} (\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s))^2 - \frac{|S^j(K)|_K}{|S^j(1)|} \sum_{s, s-1 \in S^j(1)} (\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s))^2,$$

$$J_2 = \frac{1}{K} \sum_{s, s-K \in S^j(K)} \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s) \boldsymbol{\epsilon}(s-K)^{\mathrm{T}} \mathbf{p}_{ij} - \frac{|S^j(K)|_K}{|S^j(1)|} \sum_{s, s-1 \in S^j(1)} \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s) \boldsymbol{\epsilon}(s-1)^{\mathrm{T}} \mathbf{p}_{ij},$$

$$J_3 = \frac{1}{K} \sum_{s, s-K \in S^j(K)} (\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s-K))^2 - \frac{|S^j(K)|_K}{|S^j(1)|} \sum_{s, s-1 \in S^j(1)} (\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s-1))^2.$$

With Assumption (E1) to (E3), writing $g_{m,s}^{ij} = \mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s) \boldsymbol{\epsilon}(s-m)^{\mathrm{T}} \mathbf{p}_{ij}$, by Lemma 2.7 of Bai and Silverstein (1998),

$$E\left\{ \left(\frac{1}{m} \sum_{s, s-m \in S^j(m)} g_{m,s}^{ij} \right)^2 \middle| \mathcal{F}_{-j} \right\} = O(m^{-2} n \cdot 1 + n^{-1} \cdot m^{-2} n^2 \cdot 1) = O(m^{-2} n), \text{ hence}$$

$$\frac{1}{m} \sum_{s, s-m \in S^j(m)} g_{m,s}^{ij} = O_P(m^{-1} n^{1/2}),$$

which implies that

$$J_2 = O_P(K^{-1} n^{1/2}) = O_P(n^{-1/6}).$$

We can further decompose $J_1 = J_{11} - J_{12} + J_{13}$, where

$$J_{11} = \frac{1}{K} \sum_{s, s-K \in S^j(K)} ((\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s))^2 - \mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}_{\epsilon,s}^j \mathbf{p}_{ij}),$$

$$J_{12} = \frac{|S^j(K)|_K}{|S^j(1)|} \sum_{s, s-1 \in S^j(1)} ((\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\epsilon}(s))^2 - \mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}_{\epsilon,s}^j \mathbf{p}_{ij}),$$

$$J_{13} = \frac{1}{K} \sum_{s, s-K \in S^j(K)} \mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}_{\epsilon,s}^j \mathbf{p}_{ij} - \frac{|S^j(K)|_K}{|S^j(1)|} \sum_{s, s-1 \in S^j(1)} \mathbf{p}_{ij}^{\mathrm{T}} \mathbf{\Sigma}_{\epsilon,s}^j \mathbf{p}_{ij}.$$

Consider

$$
\begin{aligned}
J_{13} &= \frac{1}{K}\sum_{s,s-K\in S^j(K)} \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^{j}\mathbf{p}_{ij} - \frac{1}{K}\sum_{s,s-1\in S^j(1)} \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^{j}\mathbf{p}_{ij} + \frac{K-1}{Kn(j)}\sum_{s,s-1\in S^j(1)} \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}\mathbf{p}_{ij} \\
&= -\frac{1}{K}\sum_{s=1}^{K-1} \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^{j}\mathbf{p}_{ij} + \frac{K-1}{Kn(j)}\sum_{s=1}^{n(j)} \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^{j}\mathbf{p}_{ij} \\
&= \left(\mathbf{p}_{ij}^{\mathrm{T}}E(\boldsymbol{\Sigma}_{\epsilon,s}^{j})\mathbf{p}_{ij} - \frac{1}{K}\sum_{s=1}^{K-1} \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^{j}\mathbf{p}_{ij}\right) + \left(\frac{1}{n(j)}\sum_{s=1}^{n(j)} \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^{j}\mathbf{p}_{ij} - \mathbf{p}_{ij}^{\mathrm{T}}E(\boldsymbol{\Sigma}_{\epsilon,s}^{j})\mathbf{p}_{ij}\right) \\
&- \frac{1}{Kn(j)}\sum_{s=1}^{n(j)} \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^{j}\mathbf{p}_{ij} \\
&= O_P(K^{-1/2}) + O_P(n^{-1/2}) + O_P(K^{-1}) = O_P(n^{-1/3}),
\end{aligned}
$$

where the last line used the weak law of large number given $\mathcal{F}_{-j}$.

Now define $g_s^{ij} = \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\epsilon}(s) - \mathbf{p}_{ij}^{\mathrm{T}}\boldsymbol{\Sigma}_{\epsilon,s}^{j}\mathbf{p}_{ij}$. Using Lemma 2.7 of Bai and Silverstein (1998) under Assumption (E1) to (E3), we have

$$
\begin{aligned}
E(J_{11}^2|\mathcal{F}_{-j}\cup\{\boldsymbol{\Sigma}_{\epsilon,u}, u\in[0,1]\}) &= K^{-2}\sum_{s,s-K\in S^j(K)} E((g_s^{ij})^2|\mathcal{F}_{-j}\cup\{\boldsymbol{\Sigma}_{\epsilon,u}, u\in[0,1]\}) \\
&+ K^{-2}\sum_{s_1\neq s_2} E(g_{s_1}^{ij}g_{s_2}^{ij}|\mathcal{F}_{-j}\cup\{\boldsymbol{\Sigma}_{\epsilon,u}, u\in[0,1]\}) \\
&= O(K^{-2}n\cdot 1 + K^{-2}n^2\cdot n^{-1}\cdot 1) = O(n^{-1/3}).
\end{aligned}
$$

The above implies that

$$J_{11} = O_P(n^{-1/6}) = J_{12}.$$

The rates for J_{11}, J_{12} and J_{13} imply that

$$J_1 = O_P(n^{-1/6}) = J_3,$$

so that combining with the rate of J_2, we have

$$I_{3,1}(K) - \frac{|S^j(K)|_K}{|S^j(1)|}I_{3,1}(1) = O_P(n^{-1/6}L).$$

Finally, among $I_{3,2}$ to $I_{3,6}$, when $m = K \asymp n^{2/3}$, the dominating term is $I_{3,5} = O_P(n^{-1/4}L^{3/4})$, while it is $I_{3,2} = O_P(p^{1/3})$ when $m = 1$. Hence

$$\max_{\substack{i=1,\dots,p\\ j=1,\dots,L}} \left| \frac{I_3}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}} \right| = O_P(n^{-1/6}L) + O_P(n^{-1/4}L^{3/4}) + O_P(K^{-1} \cdot p^{1/3})$$
$$= O_P(n^{-1/6}L) = O_P(n^{-1/6}),$$

since L is finite. This completes the proof of the lemma. □

Proof of Theorem 4.1, 4.2. Combining the results of Lemma 4.1, 4.3 and 4.4, we have

$$\max_{\substack{i=1,\dots,p\\ j=1,\dots,L}} \left| \frac{\mathbf{p}_{ij}^{\mathrm{T}} \widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}} - 1 \right| \leq \max_{\substack{i=1,\dots,p\\ j=1,\dots,L}} \left| \frac{I_1}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}} - 1 \right| + 2 \max_{\substack{i=1,\dots,p\\ j=1,\dots,L}} \left| \frac{I_2}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}} \right|$$
$$+ \max_{\substack{i=1,\dots,p\\ j=1,\dots,L}} \left| \frac{I_3}{\mathbf{p}_{ij}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{p}_{ij}} \right| = O_P(n^{-1/6}).$$

Note that the above result is equivalent to the main result in Theorem 4.1. For the second main result,

$$\|\widehat{\boldsymbol{\Sigma}}(0,1)\boldsymbol{\Sigma}_{\text{Ideal}}^{-1} - \mathbf{I}_p\| = \left\| \sum_{j=1}^{L} (\widehat{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)\boldsymbol{\Sigma}_{\text{Ideal}}(\tau_{j-1}, \tau_j)^{-1} - \mathbf{I}_p)\boldsymbol{\Sigma}_{\text{Ideal}}(\tau_{j-1}, \tau_j)\boldsymbol{\Sigma}_{\text{Ideal}}(0,1)^{-1} \right\|$$
$$\leq \sum_{j=1}^{L} \|\widehat{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j)\boldsymbol{\Sigma}_{\text{Ideal}}(\tau_{j-1}, \tau_j)^{-1} - \mathbf{I}_p\|$$
$$\cdot \left\| \operatorname{diag}(\mathbf{P}_{-j}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{P}_{-j}) \cdot \left(\operatorname{diag}(\mathbf{P}_{-j}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{P}_{-j}) + \sum_{i \neq j} \mathbf{P}_{-j}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{i-1}, \tau_i)\mathbf{P}_{-j} \right)^{-1} \right\|$$
$$= O_P\left(Ln^{-1/6} \cdot \max_{j=1,\dots,L} \left\| \left(\mathbf{I}_p + \sum_{i \neq j} \mathbf{P}_{-j}^{\mathrm{T}} \boldsymbol{\Sigma}_{\text{Ideal}}(\tau_{i-1}, \tau_i)\mathbf{P}_{-j} \operatorname{diag}^{-1}(\mathbf{P}_{-j}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j)\mathbf{P}_{-j}) \right)^{-1} \right\| \right)$$
$$= O_P(n^{-1/6}).$$

Hence these completes the proof of Theorem 4.1 and the equivalent part in Theorem 4.2 under jumps removed data.

To complete the proof of Theorem 4.2, note that for a generic constant $C > 0$,

$$\begin{aligned}\Big\| \sum_{0\le t\le 1} (\Delta \mathbf{J}_t \Delta \mathbf{J}_t^{\mathrm{T}} - \Delta \widehat{\mathbf{J}}_t \Delta \widehat{\mathbf{J}}_t^{\mathrm{T}}) \Big\| &\le C \max_{0\le t\le 1} \|\Delta \mathbf{J}_t \Delta \mathbf{J}_t^{\mathrm{T}} - \Delta \widehat{\mathbf{J}}_t \Delta \widehat{\mathbf{J}}_t^{\mathrm{T}})\| \\ &\le 2C \max_{0\le t\le 1} \|\Delta \mathbf{J}_t - \Delta \widehat{\mathbf{J}}_t\| \cdot \|\Delta \mathbf{J}_t\| + C \max_{0\le t\le 1} \|\Delta \mathbf{J}_t - \Delta \widehat{\mathbf{J}}_t\|^2 \\ &= O_P(n^{-1/4} L^{-1/4}),\end{aligned}$$

where the first line used Assumption (W2) that there are only finite number of jumps in $[0, 1]$ for each stock, and the second line used Assumption (W3) that there are only finite number of cojumps, with rate of jumps removal given as in Fan and Wang (2007). This completes the proof of Theorem 4.2. □

Proof of Theorem 4.3. Define $\mathbf{D}_j = \mathrm{diag}(\mathbf{P}_{-j}^{\mathrm{T}} \boldsymbol{\Sigma}(\tau_{j-1}, \tau_j) \mathbf{P}_{-j})$ and $\widetilde{\mathbf{D}}_j = \mathrm{diag}(\mathbf{P}_{-j}^{\mathrm{T}} \widetilde{\boldsymbol{\Sigma}}(\tau_{j-1}, \tau_j) \mathbf{P}_{-j})$. Define $\mathbf{e}_i$ to be the unit vector with 1 on the ith position and 0 elsewhere, and $\|\mathbf{A}\|_1 = \max_j \sum_i |a_{ij}|$ the L_1 norm of a matrix $\mathbf{A}$. Then for some $i = 1, \ldots, p$,

$$\begin{aligned}p^{1/2} \|\widehat{\mathbf{w}}_{\mathrm{opt}}\|_{\max} &= \frac{p^{1/2} |\mathbf{e}_i^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}(0,1)^{-1} \mathbf{1}_p|}{\mathbf{1}_p^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}(0,1)^{-1} \mathbf{1}_p} \le \frac{p^{1/2} \|\widehat{\boldsymbol{\Sigma}}(0,1)^{-1}\|_1}{p \lambda_{\min}(\widehat{\boldsymbol{\Sigma}}(0,1)^{-1})} \le \frac{p^{1/2} \cdot p^{1/2} / \lambda_{\min}(\widehat{\boldsymbol{\Sigma}}(0,1))}{p / \lambda_{\max}(\widehat{\boldsymbol{\Sigma}}(0,1))} \\ &\le \frac{\sum_{j=1}^L \lambda_{\max}(\widetilde{\mathbf{D}}_j)}{\sum_{j=1}^L \lambda_{\min}(\widetilde{\mathbf{D}}_j)} \\ &\le \frac{L \max_{1\le j\le L} \lambda_{\max}(\widetilde{\mathbf{D}}_j \mathbf{D}_j^{-1} - \mathbf{I}_p) \lambda_{\max}(\mathbf{D}_j) + \sum_{j=1}^L \lambda_{\max}(\mathbf{D}_j)}{L \min_{1\le j\le L} \lambda_{\min}(\widetilde{\mathbf{D}}_j \mathbf{D}_j^{-1} - \mathbf{I}_p) \lambda_{\min}(\mathbf{D}_j) + \sum_{j=1}^L \lambda_{\min}(\mathbf{D}_j)} \\ &\le \frac{(\max_{1\le j\le L} \lambda_{\max}(\widetilde{\mathbf{D}}_j \mathbf{D}_j^{-1} - \mathbf{I}_p) + 1) \max_{1\le j\le L} \lambda_{\max}(\mathbf{D}_j)}{(\min_{1\le j\le L} \lambda_{\min}(\widetilde{\mathbf{D}}_j \mathbf{D}_j^{-1} - \mathbf{I}_p) + 1) \min_{1\le j\le L} \lambda_{\min}(\mathbf{D}_j)} \\ &\xrightarrow{\mathrm{P}} \frac{\max_{1\le j\le L} \lambda_{\max}(\mathbf{D}_j)}{\min_{1\le j\le L} \lambda_{\min}(\mathbf{D}_j)} \le \frac{\max_{1\le j\le L} \lambda_{\max}(\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j))}{\min_{1\le j\le L} \lambda_{\min}(\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j))},\end{aligned}$$

where the last line follows from the results of Theorem 4.1 and Theorem 4.2. For the theoretical minimum-variance portfolio,

$$\begin{aligned}p^{1/2} \|\mathbf{w}_{\mathrm{theo}}\|_{\max} &= \frac{p^{1/2} |\mathbf{e}_i^{\mathrm{T}} \boldsymbol{\Sigma}(0,1)^{-1} \mathbf{1}_p|}{\mathbf{1}_p^{\mathrm{T}} \boldsymbol{\Sigma}(0,1)^{-1} \mathbf{1}_p} \le \frac{p^{1/2} \|\boldsymbol{\Sigma}(0,1)^{-1}\|_1}{p \lambda_{\min}(\boldsymbol{\Sigma}(0,1)^{-1})} \le \frac{p^{1/2} \cdot p^{1/2} / \lambda_{\min}(\boldsymbol{\Sigma}(0,1))}{p / \lambda_{\max}(\boldsymbol{\Sigma}(0,1))} \\ &\le \frac{\sum_{j=1}^L \lambda_{\max}(\mathbf{D}_j)}{\sum_{j=1}^L \lambda_{\min}(\mathbf{D}_j)} = \frac{\max_{1\le j\le L} \lambda_{\max}(\mathbf{D}_j)}{\min_{1\le j\le L} \lambda_{\min}(\mathbf{D}_j)} \le \frac{\max_{1\le j\le L} \lambda_{\max}(\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j))}{\min_{1\le j\le L} \lambda_{\min}(\boldsymbol{\Sigma}(\tau_{j-1}, \tau_j))}.\end{aligned}$$

For the actual risk bound, define $\mathbf{R} = \sum_{j=1}^{L} \mathbf{P}_{-j}(\widetilde{\mathbf{D}}_j\mathbf{D}_j^{-1} - \mathbf{I}_p)\mathbf{D}_j\mathbf{P}_{-j}^{\mathrm{T}}$. We first consider the case of no pervasive factors. Consider

$$\begin{aligned}\widehat{\mathbf{\Sigma}}(0,1)^{-1} &= \left(\sum_{j=1}^{L}\mathbf{P}_{-j}\widetilde{\mathbf{D}}_j\mathbf{P}_{-j}^{\mathrm{T}}\right)^{-1} = \left(\sum_{j=1}^{L}\mathbf{P}_{-j}(\widetilde{\mathbf{D}}_j\mathbf{D}_j^{-1} - \mathbf{I}_p)\mathbf{D}_j\mathbf{P}_{-j}^{\mathrm{T}} + \sum_{j=1}^{L}\mathbf{P}_{-j}\mathbf{D}_j\mathbf{P}_{-j}^{\mathrm{T}}\right)^{-1} \\ &= (\mathbf{I}_p + \mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1}\mathbf{R})^{-1}\mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1} \\ &= \mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1} + \sum_{k\geq 1}\Big(-\mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1}\mathbf{R}\Big)^k\mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1},\end{aligned}$$

where the Neumann's series expansion in the last line is valid since

$$\begin{aligned}\sum_{k\geq 0}\|\mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1}\|^k\|\mathbf{R}\|^k &\leq 1 + \sum_{k\geq 1}\frac{\|\mathbf{R}\|^k}{\lambda_{\min}^k(\mathbf{\Sigma}_{\text{Ideal}}(0,1))} \\ &\leq 1 + \sum_{k\geq 1}\frac{L^k\max_{1\leq j\leq L}\|\widetilde{\mathbf{D}}_j\mathbf{D}_j^{-1} - \mathbf{I}_p\|^k\max_{1\leq j\leq L}\|\mathbf{\Sigma}(\tau_{j-1},\tau_j)\|^k}{L^k\min_{1\leq j\leq L}\lambda_{\min}^k(\mathbf{\Sigma}(\tau_{j-1},\tau_j))} \\ &\xrightarrow{\mathbf{P}} 1 < \infty,\end{aligned}$$

where the last line follows from the results in Theorem 4.1 and 4.2. This implies that, in probability,

$$\|\widehat{\mathbf{\Sigma}}(0,1)^{-1} - \mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1}\| \leq \lambda_{\max}(\mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1})\sum_{k\geq 1}\frac{\|\mathbf{R}\|^k}{\lambda_{\min}^k(\mathbf{\Sigma}_{\text{Ideal}}(0,1))} \xrightarrow{\mathbf{P}} 0. \tag{4.43}$$

With the above, consider the decomposition $pR(\widehat{\mathbf{w}}_{\text{opt}}) = I_1 + I_2 + I_3$, where

$$\begin{aligned}I_1 &= \frac{p\mathbf{1}_p^{\mathrm{T}}(\widehat{\mathbf{\Sigma}}(0,1)^{-1} - \mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1})\mathbf{\Sigma}(0,1)\widehat{\mathbf{\Sigma}}(0,1)^{-1}\mathbf{1}_p}{(\mathbf{1}_p^{\mathrm{T}}\widehat{\mathbf{\Sigma}}(0,1)^{-1}\mathbf{1}_p)^2}, \\ I_2 &= \frac{p\mathbf{1}_p^{\mathrm{T}}\mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1}\mathbf{\Sigma}(0,1)(\widehat{\mathbf{\Sigma}}(0,1)^{-1} - \mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1})\mathbf{1}_p}{(\mathbf{1}_p^{\mathrm{T}}\widehat{\mathbf{\Sigma}}(0,1)^{-1}\mathbf{1}_p)^2}, \\ I_3 &= \frac{p\mathbf{1}_p^{\mathrm{T}}\mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1}\mathbf{\Sigma}(0,1)\mathbf{\Sigma}_{\text{Ideal}}(0,1)^{-1}\mathbf{1}_p}{(\mathbf{1}_p^{\mathrm{T}}\widehat{\mathbf{\Sigma}}(0,1)^{-1}\mathbf{1}_p)^2}.\end{aligned}$$

By (4.43), with $\|\boldsymbol{\Sigma}(0,1)\| \leq C$ where C is a generic constant since there are no pervasive factors,

$$|I_1| \leq \frac{p^2\|\widehat{\boldsymbol{\Sigma}}(0,1)^{-1} - \boldsymbol{\Sigma}_{\text{Ideal}}(0,1)^{-1}\| \cdot C \cdot (\|\widehat{\boldsymbol{\Sigma}}(0,1)^{-1} - \boldsymbol{\Sigma}_{\text{Ideal}}(0,1)^{-1}\| + \lambda_{\max}(\boldsymbol{\Sigma}_{\text{Ideal}}(0,1)^{-1}))}{p^2\Big(\lambda_{\min}(\boldsymbol{\Sigma}_{\text{Ideal}}(0,1)^{-1}) - \|\widehat{\boldsymbol{\Sigma}}(0,1)^{-1} - \boldsymbol{\Sigma}_{\text{Ideal}}(0,1)^{-1}\|\Big)^2}$$
$$\xrightarrow{\mathbf{P}} 0.$$

Similarly, $|I_2| \xrightarrow{\mathbf{P}} 0$. For I_3, by (4.43),

$$\begin{aligned}
|I_3| &\leq \frac{p^2\lambda_{\max}^2(\boldsymbol{\Sigma}_{\text{Ideal}}(0,1)^{-1})\lambda_{\max}(\boldsymbol{\Sigma}(0,1))}{p^2\Big(\lambda_{\min}(\boldsymbol{\Sigma}_{\text{Ideal}}(0,1)^{-1}) - \|\widehat{\boldsymbol{\Sigma}}(0,1)^{-1} - \boldsymbol{\Sigma}_{\text{Ideal}}(0,1)^{-1}\|\Big)^2} \\
&\xrightarrow{\mathbf{P}} \frac{\lambda_{\max}^2(\boldsymbol{\Sigma}_{\text{Ideal}}(0,1))}{\lambda_{\min}^2(\boldsymbol{\Sigma}_{\text{Ideal}}(0,1))}\lambda_{\max}(\boldsymbol{\Sigma}(0,1)) \\
&\leq \left(\frac{\sum_{j=1}^L \lambda_{\max}(\boldsymbol{\Sigma}(\tau_{j-1},\tau_j))}{\sum_{j=1}^L \lambda_{\min}(\boldsymbol{\Sigma}(\tau_{j-1},\tau_j))}\right)^2 \lambda_{\max}(\boldsymbol{\Sigma}(0,1)) \\
&= \left(\frac{\max_{1\leq j\leq L} \lambda_{\max}(\boldsymbol{\Sigma}(\tau_{j-1},\tau_j))}{\min_{1\leq j\leq L} \lambda_{\min}(\boldsymbol{\Sigma}(\tau_{j-1},\tau_j))}\right)^2 \lambda_{\max}(\boldsymbol{\Sigma}(0,1)),
\end{aligned}$$

which leads to the result in the theorem.

If there are pervasive factors, abbreviating $\boldsymbol{\Sigma}(0,1)$ as $\boldsymbol{\Sigma}$ etc, consider

$$\begin{aligned}
R(\widehat{\mathbf{w}}_{\text{opt}}) &= \frac{\mathbf{1}_p^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}_p}{(\mathbf{1}_p^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}_p)^2} \leq \frac{\lambda_{\max}(\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma})}{\mathbf{1}_p^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}_p} \leq \frac{\lambda_{\max}(\widehat{\boldsymbol{\Sigma}})\lambda_{\max}(\boldsymbol{\Sigma})}{p\lambda_{\min}(\widehat{\boldsymbol{\Sigma}})} \\
&= O_P(\lambda_{\max}(\boldsymbol{\Sigma})),
\end{aligned}$$

where the last line follows from the results in Theorem 4.1 and 4.2. For the actual risk bound for $\mathbf{w}_{\text{theo}}$,

$$pR(\mathbf{w}_{\text{theo}}) = \frac{p}{\mathbf{1}_p^{\mathrm{T}}\boldsymbol{\Sigma}(0,1)^{-1}\mathbf{1}_p} \leq \lambda_{\max}(\boldsymbol{\Sigma}(0,1)).$$

This completes the proof of the theorem. □

Chapter 5

Nonlinear Shrinkage Estimation of Large Integrated Covariance Matrices

Declaration This chapter is based on joint work with Dr. Clifford Lam and Dr. Charlie Hu as published by Biometrika (Lam et al., 2017). Under university policy, this chapter only include the joint work with Dr. Clifford Lam.

5.1 Introduction

Intraday data on financial asset returns are of increasing interest for portfolio allocation and risk management (Fan et al., 2012). Models for such data need to account for rapid changes in volatility during a trading day. To capture such changes, it is natural to consider covolatility processes and to combine covariances between pairs of asset returns over time through what is called an integrated covariance matrix. There are various challenges in estimating this matrix (Aït-Sahalia et al., 2005; Asparouhova et al., 2013).

Similar to Chapter 4, we consider the bias that arises when the number of assets p is large. Specifically, we suppose that p has the same order as the sample size n, i.e., $p/n \to c > 0$ for some constant $c > 0$. If synchronous time data points are observed, a natural estimator of the integrated covariance matrix can be obtained from an empirical covariance matrix of the observed returns. However, this estimator suffers

from bias, which can be expressed in terms of the bias of its extreme eigenvalues (Bai and Silverstein, 2010).

To rectify this bias problem, many researchers have focused on regularized estimation of covariance or precision matrices with special structures, such as banded (Bickel and Levina, 2008b) or sparse covariance matrix (Bickel and Levina, 2008a; Cai and Zhou, 2012; Lam and Fan, 2009; Rothman et al., 2008), sparse precision matrix (Friedman et al., 2008; Meinshausen and Bühlmann, 2006), a spiked covariance matrix from a factor model (Fan et al., 2008, 2011), or combinations of these (Fan et al., 2013). Instead of assuming a particular structrue for the true covariance matrix, nonlinear shrinkage of eigenvalues (Lam, 2016; Ledoit and Wolf, 2012) are also well researched.

In this chapter, we modify the method proposed in Lam (2016) to achieve nonlinear shrinkage of eigenvalues in a covariance matrix. Different from Chapter 4, we do not consider microstructure noise in this chapter, although this proposed estimator share very similar settings as the previous chapter. Our method produces a positive definite estimator of the integrated covariance matrix asymptotically almost surely, and involves only eigendecompositions of matrices of size $p \times p$, which are not computationally expensive when p is of the order of hundreds, the typical order in portfolio allocation. We also present the maximum exposure and actual risk bounds for minimum variance portfolio construction using our estimator. The maximum exposure bound is of particular importance, as it is shared by the theoretical minimum-variance portfolio which assumes that the integrated covariance matrix is known.

The rest of the chapter is organized as follows. We first present the framework for the data together with the notations and the main assumptions to be used in Chapter 5.2, with our proposed estimator is presented in Chapter 5.2.3. Chapter 5.3 presents all related theories. Simulation results and a real data example of portfolio allocation is presented in Chapter 5.4. All proofs are given in the supplementary materials (Lam et al., 2017) available at Biometrika website (https://doi.org/10.1093/biomet/asx021), which is not part of this thesis.

5.2 Framework and Methodology

5.2.1 Integrated and Realized Covariance Matrices

Let $\mathbf{X}_t = (X_t^{(1)}, \cdots, X_t^{(p)})^{\mathrm{T}}$ be a p-dimensional log-price diffusion process modeled by

$$d\mathbf{X}_t = \boldsymbol{\mu}_t dt + \boldsymbol{\Theta}_t d\mathbf{W}_t, \ t \in [0,1], \tag{5.1}$$

where $\boldsymbol{\mu}_t$ is the drift, $\boldsymbol{\Theta}_t$ is a $p \times p$ matrix of instantaneous covolatility process, and $\mathbf{W}_t = (W_t^{(1)}, \ldots, W_t^{(p)})^{\mathrm{T}}$ is a p-dimensional standard Brownian motion. We want to estimate the integrated covariance matrix, defined by

$$\boldsymbol{\Sigma}_0 = \int_0^1 \boldsymbol{\Theta}_t \boldsymbol{\Theta}_t^{\mathrm{T}} dt.$$

This matrix is important in risk management, the hedging and pricing of financial derivatives, and portfolio allocation, to name but a few areas of finance (Hounyo, 2017). In portfolio allocation, $\boldsymbol{\Sigma}_0$ replaces the usual population covariance matrix for intraday data. If $\boldsymbol{\Theta}_t$ is constant, then we can take $\boldsymbol{\Theta}_t = \boldsymbol{\Sigma}_0^{1/2}$, and $\boldsymbol{\Sigma}_0$ is just the usual covariance matrix for asset returns.

In this chapter, we consider sparsely sampled return data synchronized by refresh times (Andersen et al., 2001; Barndorff-Nielsen et al., 2011). Suppose that we observe $\mathbf{X}_t$ at synchronous time points $\tau_{n,\ell}, \ell = 0, \ldots, n$. The realized covariance matrix is then

$$\boldsymbol{\Sigma}_p^{\mathrm{RCV}} = \sum_{\ell=1}^{n} \Delta\mathbf{X}_\ell \Delta\mathbf{X}_\ell^{\mathrm{T}}, \quad \text{where } \Delta\mathbf{X}_\ell = \mathbf{X}_{\tau_{n,\ell}} - \mathbf{X}_{\tau_{n,\ell-1}}.$$

Jacod and Protter (1998) shows that as $n \to \infty, \boldsymbol{\Sigma}_p^{\mathrm{RCV}} \to \boldsymbol{\Sigma}_0$ weakly when p is fixed.

5.2.2 Time Variation Adjusted Realized Covariance Matrix

In this section, we present the important contributions from Zheng and Li (2011). Write $dX_t^{(j)} = \mu_t^{(j)} dt + \sigma_t^{(j)} dZ_t^{(j)}, j = 1, \ldots, p$, where $\mu_t^{(j)}, \sigma_t^{(j)}$ are assumed to be càdlàg over $[0,1]$, and the $Z_t^{(j)}$'s are one dimensional standard Brownian motions. Define $\langle X, Y \rangle_t$ to be the quadratic covariation between the processes X and Y.

(S1) The correlation matrix process of $\mathbf{Z}_t = (Z_t^{(1)}, \ldots, Z_t^{(p)})^{\mathrm{T}}$, $\langle Z^{(j)}, Z^{(k)} \rangle_t / t, 1 \leq j, k \leq p$, is constant and non-zero on $(0,1]$ for each pair of j, k. Further-

more, the correlation matrix process of $\mathbf{X}_t$, $\int_0^t \sigma_s^{(j)}\sigma_s^{(k)} d\langle Z^{(j)}, Z^{(k)}\rangle_s \{\int_0^t (\sigma_s^{(j)})^2 ds \cdot \int_0^t (\sigma_s^{(k)})^2 ds\}^{-1/2}, 1 \le j, k \le p$, is constant on $(0, 1]$ for each pair of j, k.

Then, by Proposition 4 in Zheng and Li (2011), there exists a càdlàg process $(\gamma_t)_{t\in[0,1]}$ and a $p \times p$ matrix $\mathbf{\Lambda}$ satisfying $\mathrm{tr}(\mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}) = p$ such that we can make the decomposition $\mathbf{\Theta}_t = \gamma_t \mathbf{\Lambda}$ implying that $\mathbf{\Sigma}_0 = (\int_0^1 \gamma_t^2 dt)\mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}$. The time-variation adjusted realized covariance matrix is defined by

$$\check{\mathbf{\Sigma}}_p = \frac{\mathrm{tr}(\mathbf{\Sigma}_p^{\mathrm{RCV}})}{p}\check{\mathbf{\Phi}}, \quad \text{where } \check{\mathbf{\Phi}} = \frac{p}{n}\sum_{\ell=1}^{n} \frac{\Delta\mathbf{X}_\ell \Delta\mathbf{X}_\ell^{\mathrm{T}}}{\|\Delta\mathbf{X}_\ell\|^2}, \tag{5.2}$$

and $\|\cdot\|$ denotes the L_2 norm of a vector. It is shown in Zheng and Li (2011) that $\mathrm{tr}(\mathbf{\Sigma}_p^{\mathrm{RCV}})/p$ is a good estimator for $\int_0^1 \gamma_t^2 dt$, while $\check{\mathbf{\Phi}}$ estimates $\mathbf{\Phi} = \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}$.

5.2.3 Nonlinear Shrinkage Estimator

The estimator $\check{\mathbf{\Phi}}$ is a sample covariance matrix of $\mathbf{r}_\ell = p^{1/2}\Delta\mathbf{X}_\ell/\|\Delta\mathbf{X}_\ell\|, \ell = 1, \ldots, n$, the self-normalized returns. Under the setting $p/n \to c > 0$, the eigenvalues in $\check{\mathbf{\Phi}}$ are biased estimators of those in $\mathbf{\Phi}$. The way in which each r_ℓ is defined means that we cannot apply the nonlinear shrinkage formula of Ledoit and Wolf (2012) directly. Instead, we use the data splitting idea for nonlinear shrinkage of eigenvalues from Lam (2016).

To this end, we permute the return data M times follow Lam (2016). At the jth permutation, we split the data $\Delta\mathbf{X}^{(j)}$ into two independent parts, say $\Delta\mathbf{X}^{(j)} = (\Delta\mathbf{X}_1^{(j)}, \Delta\mathbf{X}_2^{(j)}), j = 1, \ldots, M$, with $\Delta\mathbf{X}_i^{(j)}$ having size $p \times n_i, i = 1, 2$, such that $n_1 = m$ and $n_2 = n - m$. Define $\widetilde{\mathbf{\Phi}}_i^{(j)} = n_i^{-1}\sum_{\ell\in I_{i,j}} \mathbf{r}_\ell \mathbf{r}_\ell^{\mathrm{T}}$, where $I_{i,j} = \{\ell : \Delta\mathbf{X}_\ell \in \Delta\mathbf{X}_i^{(j)}\}, i = 1, 2, j = 1, \ldots, M$. Carrying out an eigen-analysis on $\widetilde{\mathbf{\Phi}}_1^{(j)}$, suppose that $\widetilde{\mathbf{\Phi}}_1^{(j)} = \mathbf{P}_1^{(j)}\mathbf{D}_1^{(j)}\mathbf{P}_1^{(j)}$. Then we define our estimator as

$$\widehat{\mathbf{\Sigma}}_{m,M} = \frac{\mathrm{tr}(\mathbf{\Sigma}_p^{\mathrm{RCV}})}{p}\frac{1}{M}\sum_{j=1}^{M}\widehat{\mathbf{\Phi}}^{(j)}, \quad \text{where } \widehat{\mathbf{\Phi}}^{(j)} = \mathbf{P}_1^{(j)}\mathrm{diag}(\mathbf{P}_1^{(j)\mathrm{T}}\widetilde{\mathbf{\Phi}}_2^{(j)}\mathbf{P}_1^{(j)})\mathbf{P}_1^{(j)\mathrm{T}}, \tag{5.3}$$

where $\mathrm{diag}(\cdot)$ sets all non-diagonal elements of a matrix to zero. Compared to the estimator $\check{\mathbf{\Sigma}}_p$ in (5.2), we substitute the sample covariance matrix $\check{\mathbf{\Phi}}$ to the averaged NERCOME estimator $\widehat{\mathbf{\Sigma}}_{m,M}$ in (5.3). The estimator $\widehat{\mathbf{\Phi}}^{(j)}$ belongs to a class of rotation equivariant estimators $\mathbf{\Phi}(\mathbf{D}) = \mathbf{P}_1^{(j)}\mathbf{D}\mathbf{P}_1^{(j)\mathrm{T}}$, where $\mathbf{D}$ is diagonal. We

choose $\mathbf{D} = \text{diag}(\mathbf{P}_1^{(j)\mathrm{T}}\widetilde{\mathbf{\Phi}}_2^{(j)}\mathbf{P}_1^{(j)})$ solves $\min_{\mathbf{D}} \|\mathbf{P}_1^{(j)}\mathbf{D}\mathbf{P}_1^{(j)\mathrm{T}} - \mathbf{\Phi}\|_F$ and by Lemma 5.1, $\mathbf{D}^{(j)} = \text{diag}(\mathbf{P}_1^{(j)\mathrm{T}}\widetilde{\mathbf{\Phi}}_2^{(j)}\mathbf{P}_1^{(j)})$ estimates $\text{diag}(\mathbf{P}_1^{(j)\mathrm{T}}\mathbf{\Phi}\mathbf{P}_1^{(j)})$ well. We use the Frobenius norm mainly for ease of deriving theoretical results. In Theorem 5.2 we also consider the inverse Stein loss.

5.3 Asmptotic Theory and Practical Implementation

We introduce four more assumptions needed for our results to hold.

(S2) The drift in (5.1) satisfies $\boldsymbol{\mu}_t = \mathbf{0}$ for $t \in [0,1]$, and $\mathbf{\Theta}_t$ is deterministic. All eigenvalues of $\mathbf{\Theta}_t\mathbf{\Theta}_t^{\mathrm{T}}$ are bounded uniformly between zero and infinity in $t \in [0,1]$. Also, M is finite.

(S3) The observation times $\tau_{n,\ell}$'s are independent of the log-price $\mathbf{X}_t$, and there exists a constant $C > 0$ such that for all positive integer n, $\max_{1\le\ell\le n} n(\tau_{n,\ell} - \tau_{n,\ell-1}) \le C$.

(S4) Let $v_{n,1} \ge \cdots \ge v_{n,p}$ be the p eigenvalues of $\mathbf{\Phi}$. Let $H_n(v) = p^{-1}\sum_{i=1}^{p} \mathbb{1}_{\{v_{n,i}\le v\}}$ be the empirical distribution function of the $v_{n,i}$. We assume that $H_n(v)$ converges to some non-random limit H at every point of continuity of H.

(S5) The support of H defined above is the union of a finite number of compact intervals bounded away from zero and infinity. Also, there exists a compact interval in $(0, +\infty)$ that contains the support of H_n for each n.

We set $\boldsymbol{\mu}_t = \mathbf{0}$ in Assumption (S2) to make the proofs and presentation simpler. If $\boldsymbol{\mu}_t$ is slowly varing locally, the results presented here remain valid at the expense of longer proofs. The deterministic nature of $\mathbf{\Theta}_t$ is essential to the independence of $\Delta\mathbf{X}_\ell$. The uniform bounds on the eigenvalues of $\mathbf{\Theta}_t\mathbf{\Theta}_t^{\mathrm{T}}$ are needed so that the individual volatility process for each $X_t^{(i)}$ is bounded uniformly, $\int_0^1 \gamma_t^2 dt > 0$ uniformly, and finally $\|\mathbf{\Sigma}_p\| = O(1)$ uniformly. The last two assumptions are essentially assumptions (A3) and (A4) in Lam (2016) applied to $\mathbf{\Phi}$.

Lemma 5.1 *Let Assumption (S1), (S2) and (S3) hold for the log-price process $\mathbf{X}_t$ in (5.1). If $p/n \to c > 0$ and $\sum_{n_2\ge1} pn_2^{-5} < \infty$, then $\max_{j=1,\ldots,M} \|diag(\mathbf{P}_1^{(j)\mathrm{T}}\widetilde{\mathbf{\Phi}}_2^{(j)}\mathbf{P}_1^{(j)}) \cdot diag^{-1}(\mathbf{P}_1^{(j)\mathrm{T}}\mathbf{\Phi}\mathbf{P}_1^{(j)}) - 1\| \to 0$ almost surely.*

With this result, we have the following theorem.

Theorem 5.1 *Let all the assumptions in Lemma 5.1 hold. Then* $\widehat{\mathbf{\Sigma}}_{m,M}$ *defined in (5.3) is asymptotically almost surely positive definite.*

This is an important result since $\mathbf{\Sigma}_0$ is always positive definite, which is not always the case for a realized covariance matrix, especially when $p > n$. The proof are given in the supplementary materials (Lam et al., 2017).

To present the rest of the results, we introduce a benchmark ideal estimator

$$\mathbf{\Sigma}_{\text{ideal}} = \left(\int_0^1 \gamma_t^2 dt \right) \mathbf{P} \text{diag}(\mathbf{P}^{\text{T}} \mathbf{\Phi} \mathbf{P}) \mathbf{P}^{\text{T}}.$$

This is similar to $\widehat{\mathbf{\Sigma}}_{m,M}$ defined in equation (5.3), except that $\text{tr}(\mathbf{\Sigma}_p^{\text{RCV}})/p$ is replaced by the population counterpart $\int_0^1 \gamma_t^2 dt$, while $\widehat{\mathbf{\Phi}}^{(j)}$ is replaced by $\mathbf{P}\text{diag}(\mathbf{P}^{\text{T}}\mathbf{\Phi}\mathbf{P})\mathbf{P}^{\text{T}}$, where $\mathbf{P}$ is such that $\check{\mathbf{\Phi}} = \mathbf{P}\check{\mathbf{D}}\mathbf{P}^{\text{T}}$, the eigen-decomposition of $\check{\mathbf{\Phi}}$ defined in equation (5.2). Define the efficiency loss of $\widehat{\mathbf{\Sigma}}$ as

$$EL(\mathbf{\Sigma}_0, \widehat{\mathbf{\Sigma}}) = 1 - \frac{L(\mathbf{\Sigma}_0, \mathbf{\Sigma}_{\text{Ideal}})}{L(\mathbf{\Sigma}_0, \widehat{\mathbf{\Sigma}})},$$

where $L(\mathbf{\Sigma}_0, \widehat{\mathbf{\Sigma}})$ is a loss function. We consider the Frobenius loss $L(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}) = \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_p\|_F^2$, and the inverse Stein's loss function, $L(\mathbf{\Sigma}_p, \widehat{\mathbf{\Sigma}}) = \text{tr}(\mathbf{\Sigma}_p\widehat{\mathbf{\Sigma}}^{-1}) - \log\det(\mathbf{\Sigma}_p\widehat{\mathbf{\Sigma}}^{-1}) - p$. If $\widehat{\mathbf{\Sigma}}$ incurs a larger loss than $\mathbf{\Sigma}_{\text{Ideal}}$, then $EL(\mathbf{\Sigma}_0, \widehat{\mathbf{\Sigma}}) > 0$, and vice versa.

Theorem 5.2 *Let all the assumptions in Lemma 5.1 hold, together with Assumption (S4) and (S5). Moreover, if* $n_1/n \to 1$ *and* $n_2 \to \infty$, *then* $EL(\mathbf{\Sigma}_0, \widehat{\mathbf{\Sigma}}_{m,M}) \leq 0$ *asymptotically almost surely with respect to both the Frobenius and the inverse Stein's loss functions, provided that* $p^{-1}L(\mathbf{\Sigma}_0, \mathbf{\Sigma}_{Ideal}) \nrightarrow 0$ *almost surely.*

The requirement that $p^{-1}L(\mathbf{\Sigma}_0, \mathbf{\Sigma}_{\text{Ideal}}) \nrightarrow 0$ almost surely eliminates the case where $\mathbf{\Sigma}_0 = (\int_0^1 \gamma_t^2 dt)\mathbf{I}_p$, when both loss functions attain zero for $\mathbf{\Sigma}_{\text{Ideal}}$. Simulation confrims that $\widehat{\mathbf{\Sigma}}_{m,M}$ performs well even in this special case.

To find the best split location m empirically, we minimize

$$g(m) = \left\| \frac{1}{M} \sum_{j=1}^{M} (\widehat{\mathbf{\Phi}}^{(j)} - \widetilde{\mathbf{\Phi}}_2^{(j)}) \right\|_F^2.$$

In practice, we use $M = 50$ which provides a good trade-off between computational complexity and estimation accuracy. We search the following split locations for minimizing $g(m)$:

$$m = [2n^{1/2}, 0.2n, 0.4n, 0.6n, 0.8n, n - 2.5n^{1/2}, n - 1.5n^{1/2}].$$

The location $2n^{1/2}$ is suitable for $\mathbf{\Sigma}_0 = (\int_0^1 \gamma_t^2 dt)\mathbf{I}_p$, while $n - 2.5n^{1/2}$ and $n - 1.5n^{1/2}$ satisfy the conditions $\sum_{n_2 \geq 1} p n_2^{-5} < \infty$, $n_1/n \to 1$ and $n_2 \to \infty$ needed in Theorem 5.2. We include $0.2n$ to $0.8n$ for boosting finite sample performance.

5.4 Empirical Results

5.4.1 Simulations with Varying γ_t

In this section, we compare our method to banding (Band) in Bickel and Levina (2008b), the condition number regularized estimator (CRC) proposed in Abadir et al. (2014), the nonlinear shrinkage method (NONLIN) in Ledoit and Wolf (2012), the principal orthogonal complement thresholding method (POET) in Fan et al. (2013), the graphical LASSO (GLASSO) in Friedman et al. (2008), and adaptive thresholding with the smmothly clipped absolute devition penalty (SCAD) in Fan and Li (2001). All these methods are applied to $\check{\mathbf{\Phi}}$ in equation (5.2).

We consider two scenarios for the diffusion process $\{\mathbf{X}_t\}$:

Design I: Piecewise constants. We take γ_t to be

$$\gamma_t = \begin{cases} 0.01 \times 7^{1/2}, & 0 \leq t < 1/4 \text{ or } 3/4 \leq t \leq 1, \\ 0.01, & 1/4 \leq t < 3/4. \end{cases}$$

Design II: Continuous path. We take γ_t to be

$$\gamma_t = (0.0009 + 0.0008 \cos(2\pi t))^{1/2}, \quad 0 \leq t \leq 1.$$

We assume $\mathbf{\Lambda} = (0.5^{|i-j|})_{i,j=1,\ldots,p}$ and the observation times are $\tau_{n,\ell} = \ell/n, \ell = 1, \ldots, n$. We generate $\{\mathbf{X}_t\}$ using model (5.1), obtaining $n = 200$ observations, and take $p = 100, 200$. For each design and (n, p) combination, we repeat the simulations 500 times and compare the mean Frobenius and inverse Stein's losses for the estimators.

We use 5-fold cross-validation to choose the tuning parameter for Band, and use $K = 3$ factors for POET with $\theta = 0.5$ as the thresholding parameter, the same as for SCAD. Finally, we use $\theta = 0.8$ for the tuning parameter for GLASSO. These parameters are chosen to allow the methods to have the best possible performances overall. Pre-setting these parameters also speeds up the simulations significantly.

Table 5.1 presents the simulation results. All methods perform better than the realized covariance, as expected. GLASSO is best at minimizing the Frobenius loss, while the CRC with $p = 100$, and POET with $p = 200$ are the best for the inverse Stein loss. Both our method and the CRC outperform NONLIN, which is expected since nonlinear shrinkage cannot readily be applied to self-normalized vectors. Although the way in which $\mathbf{\Lambda}$ is defined favours Band, that method had substantially larger standard deviations in all the settings.

5.4.2 Portfolio Allocation on NYSE Data

As an application in finance, we construct minimum-variance portfolios using seven different estimators compared in the previous section, except for GLASSO because of nonconvergence issues. Given an integrated covariance matrix $\mathbf{\Sigma}_0$, the minimum-variance portfolio solves $\min_{\mathbf{w}:\mathbf{w}^\top \mathbf{1}_p=1} \mathbf{w}^\top \mathbf{\Sigma}_p \mathbf{w}$, where $\mathbf{1}_p$ is a vector of p ones. The solution is

$$\mathbf{w}_{\text{opt}} = \frac{\mathbf{\Sigma}_0^{-1}\mathbf{1}_p}{\mathbf{1}_p^\top \mathbf{\Sigma}_0^{-1}\mathbf{1}_p}. \tag{5.4}$$

Before presenting the empirical results, we state a theorem concerning $\mathbf{w}_{\text{opt}}$ constructed with $\widehat{\mathbf{\Sigma}}_{m,M}$ substitude for $\mathbf{\Sigma}_0$. In what follows, we denote $\|\cdot\|_{\max}$ the maximum absolute value of a vector, and define the condition number of a positive semi-definite matrix $\mathbf{A}$ to be $\text{Cond}(\mathbf{A}) = \lambda_{\max}(\mathbf{A})/\lambda_{\min}(\mathbf{A})$.

Theorem 5.3 *Let all the assumptions in Lemma 5.1 hold. Then almost surely,*

$$\begin{aligned} p^{1/2}\|\widehat{\mathbf{w}}_{\text{opt}}\|_{\max} &\leq \text{Cond}(\mathbf{\Phi}), \quad p^{1/2}R(\widehat{\mathbf{w}}_{\text{opt}}) \leq \text{Cond}(\mathbf{\Phi})\lambda_{\max}^{1/2}(\mathbf{\Sigma}_p), \\ p^{1/2}\|\mathbf{w}_{\text{opt}}\|_{\max} &\leq \text{Cond}(\mathbf{\Phi}), \quad p^{1/2}R(\mathbf{w}_{\text{opt}}) \leq \lambda_{\max}^{1/2}(\mathbf{\Sigma}_p), \end{aligned}$$

where $\widehat{\mathbf{w}}_{\text{opt}}$ is the weight in (5.4) with $\mathbf{\Sigma}_0$ substituted by $\widehat{\mathbf{\Sigma}}_{m,M}$. The function $R(w) = (w^\top \mathbf{\Sigma}_p w)^{1/2}$ represents the actual risk when investing using w as the portfolio weights.

$n = 200$	$p = 100$				p=200			
	Design I losses		Design II losses		Design I losses		Design II losses	
	Frobenius	Inverse Stein	Frobenius	Inverse Stein	Frobenius	Inverse Stein	Frobenius	Inverse Stein
RCV	94_4	$329.7_{11.7}$	207_9	$271.5_{10.3}$	157_4	–	343_9	–
Proposed	61_3	$18.7_{0.9}$	138_7	$18.7_{0.9}$	83_3	$32.5_{1.2}$	185_6	$32.4_{1.1}$
Band	73_7	$38.1_{8.7}$	165_{14}	$38.4_{7.9}$	112_{15}	$76.0_{23.1}$	252_{34}	$75.8_{27.4}$
CRC	58_3	$10.9_{0.3}$	130_7	$10.9_{0.3}$	76_3	$27.1_{0.6}$	170_6	$27.1_{0.6}$
NONLIN	65_3	$21.5_{1.3}$	147_7	$21.6_{1.3}$	91_3	$134.9_{1032.8}$	204_7	$66.7_{240.4}$
POET	77_3	$11.3_{0.6}$	175_8	$11.4_{0.6}$	112_4	$24.8_{0.9}$	252_8	$24.9_{1.0}$
GLASSO	35_0	$32.1_{0.6}$	79_1	$32.2_{0.6}$	50_0	$64.7_{0.8}$	112_1	$64.7_{0.7}$
SCAD	60_3	$16.2_{0.9}$	135_7	$16.2_{0.9}$	88_3	$54.5_{3.8}$	197_6	$54.9_{3.9}$

Table 5.1 Mean losses for different methods, with standard errors in subscript. For the Frobenius loss, all values reported are multiplied by 10000. The realized covariance matrix is poorly conditioned when $n = p = 200$, so the inverse Stein loss does not exist.

*

RCV: realized covariance;
Band: Banding;
CRC: condition number regulized;
NONLIN: nonlinear shrinkage;
POET: principal orthogonal complement thresholding;
GLASSO: Graphical Lasso;
SCAD: adaptive thresholding with smoothly clipped absolute deviation penalty.

This theorem shows that the maximum absolute weight, which we define as the maximum exposure of the portfolio, is decaying at a rate $p^{-1/2}$, the same as that for the actual risk. This maximum exposure bound iss important, since the theoretical minimum-variance portfolio satisfies the same bound. If $\text{Cond}(\mathbf{\Phi}) = 1$, the actual risk for our portfolio can also enjoy the same upper bound as its theoretial counterpart.

We consider $p = 154$ finance stocks with large capitalization from the New York Stock Exchange (NYSE). There are 82 weeks of data, from June 2014 to the end of December 2015. We downloaded all the trades of these stocks from Wharton Research Data Services (WRDS). The raw data are high-frequency. The stocks have nonsynchronous trading times and all the log-prices are contaminated by market microstructure noise (Asparouhova et al., 2013).

We consider trades in 15-minute intervals on every trading day from 9:30 to 16:00, with each log-price being the observed one from a trade right before the end of a 15-minute interval. This results in a total of $n = 10267$ synchronized return data points. Overnight returns are not included in the calculations as overnight price jumps are usually influenced by the arrival of news, which is irrelevant to the comparison of portfolios. At the start, we invest one unit of capital using (5.4) constructed from different estimators of $\mathbf{\Sigma}_0$. We consider two-week, four-week and six-week training windows and re-evaluate portfolio weights every week. We use the annualized out-of-sample standard deviation $\widehat{\sigma}$, together with the annualized portfolio return $\widehat{\mu}$ and the Sharpe ratio $\widehat{\mu}/\widehat{\sigma}$, to gauge the performance of each method. For ℓ-week training windows and a weekly re-evaluation period, $\widehat{\mu}$ and $\widehat{\sigma}$ are defined by

$$\widehat{\mu} = 52 \times \frac{1}{30-\ell} \sum_{i=\ell+1}^{30} \mathbf{w}_i^{\mathrm{T}} \mathbf{r}_i, \quad \widehat{\sigma} = \left(52 \times \frac{1}{30-\ell} \sum_{i=ell+1}^{30} (\mathbf{w}_i^{\mathrm{T}} \mathbf{r}_i - \widehat{\mu}/52)^2 \right)^{1/2}, \ \ell = 2, 4, 6,$$

where $\mathbf{w}_i$ and $\mathbf{r}_i$ are the portfolio weights and returns respectively, for the ith week. We also report the mean and the maximum of $\|\widehat{\mathbf{w}}_{\text{opt}}\|_{\max}$ over all investment periods for the portfolios constructed with different methods.

Table 5.2 shows the results. POET and SCAD are unstable, with maximum exposures going above 200% at times, meaning that the long or short position on a single stock can be over 200%. This is not practically sound without further information on the stocks. The nonlinear shrinkage method has the smallest $\widehat{\sigma}$ in all settings, followed by our method, Band and CRC. With six-week training windows, realized covariance has the second smallest $\widehat{\sigma}$, but on average the maximum exposures are much

larger than in our method and the grand average estimator. Our method has small maximum exposures while maintaining Sharpe ratios greater than 0.7 in all settings. It has the largest Sharpe ratio when we use four-week training windows.

5.5 Conclusion

We introduce a novel nonlinear shrinkage estimator for the integrated volatility matrix using the data splitting method similar to NERIVE in Chapter 4. Different from Chapter 4, we do not consider microstructure noise here although the estimator share very similar settings as NERIVE. It produces a positive definite estimator of the integrated covariance matrix asymptotically almost surely, and involvoes only eigendecompositions of matrices of size $p \times p$ that are not computationally expensive. We also present the maximum exposure and acutal risk bounds for minimum variance portfolio construction using proposed estimator. With the numerical examples and real data from stock market trading data, we demostrated that our estimator has a favorable performance in general compared to other covariance matrix estimators.

$p = 154$	Ann. return(%)	Ann. std dev.(%)	Sharpe ratio	Max. exposure(%)	Max. of max. exposure(%)
	Weekly rebalancing with 2-week training window				
RCV	21.8	12.5	1.7	$25.3_{12.5}$	81.3
Proposed	10.2	9.4	1.1	$7.2_{1.7}$	13.6
Band	12.5	8.5	1.5	$15.9_{8.7}$	39.2
CRC	10.4	8.9	1.2	$7.4_{2.1}$	14.0
NONLIN	−0.3	8.2	0.0	$5.6_{3.5}$	14.1
POET	−3.9	11.2	−0.3	$19.9_{44.2}$	399.3
SCAD	−15.5	21.2	−0.7	$29.7_{43.6}$	326.3
	Weekly rebalancing with 4-week training window				
RCV	10.8	11.0	1.0	$20.9_{11.4}$	48.4
Proposed	13.4	9.8	1.4	$8.7_{2.7}$	17.4
Band	9.3	10.0	0.9	$17.0_{7.5}$	37.6
CRC	11.4	11.1	1.0	$8.0_{1.7}$	13.3
NONLIN	7.6	7.8	1.0	$7.7_{6.0}$	22.8
POET	1.1	11.4	0.1	$20.8_{32.9}$	235.3
SCAD	−4.3	13.7	−0.3	$27.9_{97.1}$	860.6
	Weekly rebalancing with 6-week training window				
RCV	8.7	8.8	1.0	$19.6_{11.3}$	46.0
Proposed	7.5	10.2	0.7	$10.0_{4.4}$	21.7
Band	3.7	12.0	0.3	$16.2_{7.5}$	33.6
CRC	2.9	12.2	0.2	$8.7_{2.5}$	14.8
NONLIN	6.8	7.3	0.9	$9.0_{7.4}$	26.1
POET	−9.3	14.4	−0.6	$21.5_{32.5}$	259.6
SCAD	114.9	140.7	0.8	$130.3_{959.1}$	8375.3

Table 5.2 Results of the analysis for NYSE large capitalization finance stocks (standard errors are given in subscript).

*

RCV: realized covariance;
Band: Banding;
CRC: condition number regulized;
NONLIN: nonlinear shrinkage;
POET: principal orthogonal complement thresholding;
SCAD: adaptive thresholding with smoothly clipped absolute deviation penalty.

References

Abadir, K. M., Distaso, W., and Žikeš, F. (2014). Design-free estimation of variance matrices. *Journal of Econometrics*, 181(2):165–180.

Aït-Sahalia, Y., Fan, J., and Xiu, D. (2010). High-frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association*, 105(492):1504–1517.

Aït-Sahalia, Y., Mykland, P. A., and Zhang, L. (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *The review of financial studies*, 18(2):351–416.

Aït-Sahalia, Y. and Xiu, D. (2017). Using principal component analysis to estimate a high dimensional factor model with high-frequency data. *Journal of Econometrics*, 201(2):384–399.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of financial economics*, 61(1):43–76.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455):939–967.

Asparouhova, E., Bessembinder, H., and Kalcheva, I. (2013). Noisy prices and inference regarding returns. *The Journal of Finance*, 68(2):665–714.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, Z., Miao, B., and Pan, G. (2007). On asymptotics of eigenvectors of large sample covariance matrix. *The Annals of Probability*, 35(4):1532–1572.

Bai, Z. and Silverstein, J. W. (2010). *Spectral analysis of large dimensional random matrices*, volume 20. Springer.

Bai, Z. and Yin, Y. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, pages 1275–1294.

Bai, Z.-D. and Silverstein, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Annals of probability*, pages 316–345.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2009). Realized kernels in practice: Trades and quotes. *The Econometrics Journal*, 12(3).

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2011). Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics*, 162(2):149–169.

Barndorff-Nielsen, O. E. and Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):253–280.

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, pages 2577–2604.

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684.

Cai, T. T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics*, 40(4):2014–2042.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144.

Cai, T. T. and Zhou, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389–2420.

Chen, R. Y. and Mykland, P. A. (2017). Model-free approaches to discern non-stationary microstructure noise and time-varying liquidity in high-frequency data. *Journal of Econometrics*, 200(1):79–103.

Chen, X., Xu, M., and Wu, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics*, 41(6):2994–3021.

Christensen, K., Kinnebrock, S., and Podolskij, M. (2010). Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. *Journal of Econometrics*, 159(1):116–133.

Dai, C., Lu, K., and Xiu, D. (2017). Knowing factors or factor loadings, or neither? evaluating estimators of large covariance matrices with noisy and asynchronous data.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46.

DeMiguel, V., Garlappi, L., Nogales, F. J., and Uppal, R. (2009). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, pages 157–175.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *biometrika*, pages 425–455.

Epps, T. W. (1979). Comovements in stock prices in the very short run. *Journal of the American Statistical Association*, 74(366a):291–298.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197.

Fan, J. and Kim, D. (2017). Robust high-dimensional volatility matrix estimation for high-frequency factor model. *Journal of the American Statistical Association*.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J., Li, Y., and Yu, K. (2012). Vast volatility matrix estimation using high-frequency data for portfolio selection. *Journal of the American Statistical Association*, 107(497):412–428.

Fan, J., Liao, Y., and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *Annals of statistics*, 39(6):3320.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680.

Fan, J. and Wang, Y. (2007). Multi-scale jump and volatility analysis for high-frequency financial data. *Journal of the American Statistical Association*, 102(480):1349–1362.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Gilder, D., Shackleton, M. B., and Taylor, S. J. (2014). Cojumps in stock prices: Empirical evidence. *Journal of Banking & Finance*, 40:443–459.

Griffin, J. E. and Oomen, R. C. (2011). Covariance measurement in the presence of non-synchronous trading and market microstructure noise. *Journal of Econometrics*, 160(1):58–68.

Hansen, P. R. and Lunde, A. (2006). Realized variance and market microstructure noise. *Journal of Business & Economic Statistics*, 24(2):127–161.

Hounyo, U. (2017). Bootstrapping integrated covariance matrix estimators in noisy jump–diffusion models with non-synchronous trading. *Journal of Econometrics*, 197(1):130–152.

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98.

Huang, N. and Fryzlewicz, P. (2015). Novelist estimator of large correlation and covariance matrices and their inverses. *London School of Economics and Political Science: Technical report, Department of Statistics*.

Jacod, J. and Protter, P. (1998). Asymptotic error distributions for the euler method for stochastic differential equations. *The Annals of Probability*, 26(1):267–307.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327.

Kim, D., Wang, Y., and Zou, J. (2016). Asymptotic theory for large volatility matrix estimation based on high-frequency financial data. *Stochastic Processes and their Applications*, 126(11):3527–3577.

Lam, C. (2016). Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *The Annals of Statistics*, 44(3):928–953.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37(6B):4254.

Lam, C. and Feng, P. (2018). A nonparametric eigenvalue-regularized integrated covariance matrix estimator for asset return data. *Journal of Econometrics*.

Lam, C., Feng, P., and Hu, C. (2017). Nonlinear shrinkage estimation of large integrated covariance matrices. *Biometrika*, 104(2):481–488.

Ledoit, O. and Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1-2):233–264.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060.

Ledoit, O. and Wolf, M. (2013). Optimal estimation of a large-dimensional covariance matrix under stein's loss.

Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457.

Markowitz, H. (1952). Portfolio selection. *The journal of finance*, 7(1):77–91.

Meddahi, N. (2002). A theoretical comparison between integrated and realized volatility. *Journal of Applied Econometrics*, 17(5):479–508.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462.

Rio, E. (2013). Inequalities and limit theorems for weakly dependent sequences.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1).

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Technical report, Stanford University, United States.

Tao, M., Wang, Y., and Chen, X. (2013). Fast convergence rates in estimating large volatility matrices using high-frequency financial data. *Econometric Theory*, 29(4):838–856.

Tao, M., Wang, Y., Yao, Q., and Zou, J. (2011). Large volatility matrix inference via combining low-frequency and high-frequency approaches. *Journal of the American Statistical Association*, 106(495):1025–1040.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

van de Geer, S. A. (2002). On hoeffding's inequality for dependent random variables. In *Empirical process techniques for dependent data*, pages 161–169. Springer.

Wang, Y. and Zou, J. (2010). Vast volatility matrix estimation for high-frequency financial data. *The Annals of Statistics*, 38(2):943–978.

Won, J.-H., Lim, J., Kim, S.-J., and Rajaratnam, B. (2013). Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):427–450.

Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844.

Xiu, D. (2010). Quasi-maximum likelihood estimation of volatility with high frequency data. *Journal of Econometrics*, 159(1):235–250.

Xue, Y., Gencay, R., and Fagan, S. (2014). Jump detection with wavelets for high-frequency financial time series. *Quantitative Finance*, 14(8):1427–1444.

Zhang, L. (2011). Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics*, 160(1):33–47.

Zhang, L., Mykland, P. A., and Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100(472):1394–1411.

Zheng, X. and Li, Y. (2011). On the estimation of integrated covariance matrices of high dimensional diffusion processes. *The Annals of Statistics*, 39(6):3121–3151.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.